Corpus Linguistics 2015

Abstract Book

Edited by

Federica Formato and Andrew Hardie

Lancaster: UCREL

Online edition 21st July 2015, with errata corrected

Table of contents

Plenaries

Papers

Posters

Plenaries

When an uptight register lets its hair down: The historical development of grammatical complexity features in specialist academic writing

Douglas Biber
Northern Arizona University
douglas.biber@nau.edu

Using corpus-based analyses, this talk challenges widely-held beliefs about grammatical complexity, academic writing, and linguistic change in English. It challenges stereotypes about the nature of grammatical complexity, showing that embedded phrasal structures are as important as embedded dependent clauses. It challenges stereotypes about linguistic change, showing that grammatical change occurs in writing as well as speech. But perhaps most surprisingly, it challenges stereotypes about academic writing, showing that academic writing is structurally compressed (rather than elaborated); that academic writing is often *not* explicit in the expression of meaning; and that scientific academic writing has been the locus of some of the most important grammatical changes in English over the past 200 years (rather than being conservative and resistant to change).

Learner corpus research: A fast-growing interdisciplinary field

Sylviane Granger
Université catholique de Louvain
sylviane.granger@uclouvain.be

Since its emergence in the early 1990s, the field of learner corpus research (LCR) has matured and expanded significantly. In the first part of my presentation, I will provide a brief overview of the field and assess to what extent it has met the challenges which the late Geoffrey Leech identified with rare perspicacity in his preface to the first volume on learner corpora (Leech 1998). LCR has become increasingly interdisciplinary and part of my talk will be devoted to the contribution of learner corpora to the fields of corpus linguistics, second language acquisition, foreign language teaching and testing, and natural language processing. To illustrate the insights provided by LCR, I will focus on the domain of phraseology in the wide sense, i.e. the "huge area of syntagmatic prospection" (Sinclair 2004) opened up by the combined use of corpora and powerful corpus analysis techniques. Numerous learner-corpus-based studies have highlighted the difficulties that phraseological patterning represents for learners and have identified transfer from the mother tongue as a significant factor (for a survey, see Paquot & Granger 2012). Phraseological units have also been shown to be strong indicators of L2 proficiency (Crossley & Salsbury 2011; Granger & Bestgen 2014), thereby opening up innovative perspectives for language assessment and automated scoring. Combined insights from native and learner corpora can contribute to a wide range of innovative applications tailor-made to learners' attested needs. While progress in this area is very slow, the few up-and-running applications highlight the considerable potential of learner-corpus-informed resources. By way of illustration, I will describe a web-based dictionary-cum-writing aid tool focused on the phraseology of cross-disciplinary academic vocabulary (Granger & Paquot 2010 and forthcoming). This tool draws information from expert academic corpora on the typical patterning of academic words (collocations and lexical bundles) and makes use of learner corpora to identify the difficulties such patterning poses to learners. One of the most attractive features of the system is that it can be customized according to users' discipline and mother tongue background.

References

Crossley, S. & Salsbury, T.L. (2011). The development of lexical bundle accuracy and production in English second language speakers. *IRAL - International Review of Applied Linguistics in Language Teaching* 49(1), 1-26.

Granger, S. & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *IRAL - International Review of Applied Linguistics in Language Teaching* 52(3), 229-

252.

Granger, S. & Paquot, M. (2010). Customising a general EAP dictionary to meet learner needs. In Granger, S. & Paquot, M. (eds.) *eLexicography in the 21st century: New challenges, new applications.* Presses universitaires de Louvain: Louvain-la-Neuve, 87-96.

Granger, S. & Paquot, M. (forthcoming). Electronic lexicography goes local. Design and structures of a needs-driven online academic writing aid. *Lexicographica*

Leech, G. (1998). Preface: Learner corpora: what they are and what can be done with them. In Granger, S. (ed.) *Learner English on Computer.* Addison Wesley Longman: London & New York.

Paquot, M. & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32, 130-149.

Sinclair J. (2004). *Trust the Text – Language, corpus and discourse*. London: Routledge.

Exploring the interface of language and literature from a corpus linguistic point of view?

Michaela Mahlberg
University of Nottingham
michaela.mahlberg@nottingham.ac.uk

Corpus linguistics investigates language on the basis of electronically stored samples of naturally occurring texts (written or spoken). The focus on natural data emphasises the social dimension of language: the texts in a corpus are used by people in real communicative situations. So corpus linguistics can contribute to the investigation of what people do with language and how they view the world.

Literary texts create fictional worlds, but patterns in literary texts also relate to patterns that are used to talk about and make sense of the real world. This paper will explore the fuzzy boundaries between literary and non-literary texts. Corpus methods make it possible to see similarities between fictional speech and real spoken language. Corpus methods also help us discover patterns and linguistic units that are specific to the way in which narrative fiction builds textual worlds, e.g. suspensions or lexical patterns of body language presentation (Mahlberg 2013). Such literary patterns also relate to features of the real world. Importantly, the study of the nature of literary texts highlights that we need to complement corpus linguistic methods with a range of other methods and interpretative frameworks, such as psycholinguistic research (Mahlberg et al. 2014), cognitive poetics (Stockwell 2009), literary criticism and approaches in social history. Drawing on examples from Dickens's novels and other nineteenth century fiction, this paper will argue for a mixed methods approach to the study of literary texts. The paper will also illustrate some of the functionalities of the CLiC [1] tool that is being developed to support the corpus linguistic analysis of fiction as part of such an approach.

Acknowledgements

Parts of the presentation are derived from research done for the CLiC Dickens project which is supported by the UK Arts and Humanities Research Council Grant Reference AH/K005146/1.

References

Mahlberg, M. 2013. Corpus Stylistics and Dickens's *Fiction.* New York & London: Taylor & Francis.

[1] http://clic.nottingham.ac.uk

Mahlberg, M., Conklin, K. and Bisson, M.-J., 2014. "Reading Dickens's characters: employing psycholinguistic methods to investigate the cognitive reality of patterns in texts", *Language and Literature* 23(4), 369-388.

Stockwell, P. 2009. *Texture. A Cognitive Aesthetics of Reading*. Edinburgh: Edinburgh University Press.

Non-obvious meaning in CL and CADS: from 'hindsight post-dictability' to sweet serendipity*

Alan Partington
University of Bologna
alanscott.partington@unibo.it

* *serendipity*: The faculty of making happy and unexpected discoveries by accident (OED), i.e. finding out things you didn't even know you were searching for: e.g. [he] warned that readers were in danger of losing the 'serendipity' of finding a book they did not know they wanted because of the growth in online book sales (SiBol 13).

1 (Non)obviousness

In this talk I want to examine the special relevance of (non)obviousness in corpus linguistics and corpus-assisted discourse studies (CADS), drawing on case studies.

The notion that corpus linguistics can shed light on non-obvious meaning(s) and non-obvious patterns of meanings is not new; Louw (1993), Stubbs (1996) and Sinclair (2004) all allude to it. However, the concept of 'non-obvious' clearly requires some elucidation. After all, many would argue that the aim of all scientific enquiry is to uncover the non-obvious rather than the glaringly obvious, but the inductive philosophy and techniques of CL have proved exceptionally adept at serendipitous discovery. Furthermore, as Stewart (2010) reminds us, we need to consider questions like 'obvious' to whom – people have different cultural and psychological as well as language primings - and 'obvious' to what part of our mental faculties, active or passive knowledge, competence or performance, intuition or introspection. Add to this the further question - at which period of the investigation is an observation obvious or non-obvious? At the beginning or at some phase during the course of the research? Corpus linguistics in general, including the area of corpus-assisted discourse studies (CADS), is notoriously prone to the 'curse of hindsight post-dictability', that is, that once – but only once - the investigation is complete, the results either seemed rather obvious all along, or the reader is underwhelmed by the molehill of a conclusion after the mountain of research. But how much does this matter, given that corroboration is also a vital part of the scientific process? Given, though, that it remains notoriously difficult to get unsurprising findings published, how often does pessimism that we're not going to find anything

non-obvious dissuade us from undertaking a piece of research in the first place? Yet another issue is: just how much of the ability of corpus methods to uncover the non-obvious is entirely novel, and how much is it more a question of uncovering it better and/or more quickly?

The opening part will contain an overview of some of the successes of corpus linguistics in uncovering non-obviousness at the lexical-grammatical level. These include, *inter alia*: the shared meaning principle; the surprising complexity of 'small words' (*of*, *as*, *but*); collapsing the distinction between quantitative and qualitative information (probabilistic information is also functional information); the unreliability of introspection and how language description has to be recovered by processes of inference from the linguistic trace, a form of reverse engineering; the ability to track recent language change and also social, political and cultural trends (and, just as crucially, detect their absence).

I want especially to revisit evaluative (also known as *semantic* or *discourse*) prosody, one of the most striking embodiments of non-obviousness, indeed, defined by Sinclair as 'an aura of meaning which is subliminal' (Sinclair 2004: 18). Here however I want to show how the mainly bottom-up lexico-centric view of language can also *obscure* aspects of the way language functions and the way language users behave, unless it is complemented by textual and discoursal top-down perspectives, something which even CADS sometimes neglects to address.

2 Non-obviousness in CADS

The general aim of CADS is '[...] to acquaint ourselves as much as possible with the discourse type(s) in hand' (Partington, Duguid & Taylor 2013: 12), to discover by inference how discourse participants typically behave, how they typically interact and what their typical discourse aims are. But it is also interested in particular events which may stand out from this backdrop of typicality and to understand why they occurred. It therefore presents a number of different challenges from traditional CL, particularly what can corpus assistance achieve that other approaches to discourse studies struggle with? After all, they are also in the business of uncovering non-obvious meaning.

The following are some of the added values of CADS to discourse study. It can supply an overview of large numbers of texts, and by shunting between statistical analyses, close reading and analysis types half-way between the two such as concordancing, CADS is able to look at language at different *levels of abstraction*. After all, 'you cannot understand the world just by looking at it' (Stubbs 1996: 92), and abstract representations of it need to be built and then tested. Indeed, far from being unable to take context into account (the most common accusation levelled at CL), CADS contextualises, decontextualises and recontextualises language performance in a variety of ways according to research aims. Corpus techniques also greatly facilitate comparison among datasets and therefore among discourse types. They can, moreover, ensure analytical transparency and replicability (and para-replicability). And because parts of the analysis are conducted by the machine, they enable the human analyst to step outside the hermeneutic circle, to place some distance between the interpreter and the interpretation. Finally, they enable the researcher to test the validity of their observations, for instance, by searching for counterexamples ('positive cherry-picking').

With this in mind, I will examine, via reference to case studies, the various types of non-obvious meaning one can come across in CADS, which include:

- 'I knew that all along (now)' (but intuition-corroboration has an important role in science)
- 'I sensed that but didn't know why' (intuitive impressions and corpus-assisted explanations)
- 'well I never …'
- 'I never even knew I never knew that' (serendipity or 'non-obvious non-obviousness', analogous to 'unknown unknowns')
- 'it's not only non-obvious, it isn't even *there* in my corpus (and what does this mean?)'.

3 Explaining the non-obvious

Much of traditional CL has been portrayed as being largely descriptive in intent. Extracting rules of structure or the senses of lexical items from large bodies of texts does not seem to always cry out for an explanation of why these structures or senses exist in the way they do. Such a portrayal, of course, ignores the existence of CL works such as the functional explanation of the grammar of conversation contained in Biber et al (1999: 1037-1125) or Hoey's psychological explanation of language production and language system (2005). However, we do often content ourselves with the notion of language and/or speaker 'habits' without enquiring too far into the motivations behind their formation.

In discourse studies, which focus on the behaviour in particular contexts of human participants, explanations of why such behaviour might occur frequently feel absolutely necessary. And the more non-obvious a finding the more it seems to demand

an explanation. The epistemology of explanation goes back to Aristotle and beyond, but in CADS, we can focus on two types, namely inference from cause to effect and teleology, that is, inferring what the aims of a speaker or writer were in producing the text they did. An example of the former would be examining what social circumstances produced certain linguistic phenomena (numerous examples in Friginal & Hardy, 2014), whilst an example of the latter would be inferring what particular perlocutionary persuasive effects a set of participants were aiming to achieve, and who were the intended beneficiaries of their efforts (Partington 2003 on press briefings, Duguid 2007 on judicial inquiries).

Compared to the rigorous consideration paid to other parts of the research process in CADS, little explicit attention has been paid to what constitutes 'explanation', in particular, what the degree of certainty might be with which claims of having explained one's observations can be made. In fact, explanations in discourse studies are often, of necessity given the nature of the evidence (complex human interaction) well-informed speculation - speculation generated by accurate prior description - rather than hard inference. But I would argue that speculation is not a bad thing in itself, since it can act as a spur for further investigation in order to test it. And given that the evidence needed cannot always be found in the corpus material itself, it encourages the researcher to look around and outside the corpus for corroboration. I would also argue that efforts to propose alternative and competing explanations for observations are also a valuable means of testing our explanatory hypotheses (see the Conclusion to Baker et al 2013).

Finally, any natural phenomenon can be studied, represented and explained at a number of different levels 'a living organism, for example, can be studied as a collection of particles […] and as a member of a social grouping' (Honderich ed. 2005: 282). Language is somewhat analogous to Honderich's living organism, and so combining micro with macro levels of description and representation and explaining how these representations fit together and are symbiotic and mutually enhancing, as in the case mentioned earlier of evaluative prosody, is an important extension of the principle of 'total accountability' (so, not just accounting for corpus contents but also for methods of analysis) advocated by Leech (1992), the sadly missing dedicatee of this conference.

References

Baker, P., C. Gabrielatos and A. McEnery. 2013. *Discourse Analysis and Media Attitudes: The Representation of Islam in the British Press*. Cambridge: Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Duguid, A. 2007. Men at work: how those at Number 10 construct their working identity. In *Discourse, Ideology and Specialized Communication*, G. Garzone & S. Sarangi (eds), 453-484. Bern: Peter Lang.

Friginal, E. and J. Hardy 2014. *Corpus-based Sociolinguistics*. New York: Routledge.

Honderich, T. 2005. *The Oxford Companion to Philosophy*. Oxford: Oxford University Press.

Hoey, M. 2005. *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Leech, G. 1992. Corpora and theories of linguistic performance. In J. Svartvik (ed.), *Directions in CorpusLinguistics*. Proceedings of the Nobel Symposium 82, Stockholm, 4-8 August 1991, 105-122. Berlin: Mouton de Gruyter.

Louw, W. 1993. Irony in the text or insincerity in the writer? - The diagnostic potential of semantic prosodies. In M. Baker, G. Francis & E. Tognini-Bonelli (eds), *Text and Technology. In honour of John Sinclair*, 157-176. Amsterdam and Philadelphia: John Benjamins.

Partington, A. 2003. *The Linguistics of Political Argument*. Amsterdam: Benjamins.

Partington, A., A. Duguid and C. Taylor. 2013. *Patterns and Meanings in Discourse*. Amsterdam: John Benjamins.

Stewart, D. 2010. *Semantic Prosody: A Critical Evaluation.* London: Routledge.

Sinclair, J. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Stubbs, M. 1996. *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*. Oxford: Blackwell.

Papers

Semantic tagging and Early Modern collocates

Marc Alexander
University of Glasgow
marc.alexander@glasgow.ac.uk

Alistair Baron
Lancaster University
a.baron@lancaster.ac.uk

Fraser Dallachy
University of Glasgow
fraser.dallachy@glasgow.ac.uk

Scott Piao
Lancaster University
s.piao@lancaster.ac.uk

Paul Rayson
Lancaster University
p.rayson@lancaster.ac.uk

Stephen Wattam
Lancaster University
s.wattam@lancaster.ac.uk

1 Introduction

This paper describes the use of the Historical Thesaurus Semantic Tagger (HTST) to identify collocates of words in the semantic domains of *Theft* and *Authority* in the Early English Books Online (EEBO) corpus. The tagger draws its semantic classification data from the *Historical Thesaurus of English* (Kay et al. 2015). In so doing it demonstrates the value of a comprehensive and fine-grained semantic annotation system for English within corpus linguistics. Using log-likelihood measures on the semantically-annotated EEBO corpus, the paper therefore demonstrates the existence, extent, and location of significant clusters of semantic collocation in this corpus. In so doing it applies a version of Franco Moretti's 'distant reading' programme in the analysis of literary history to these texts, as well as continuing work on integrating meaning into the methodologies of corpus linguistics.

2 Word collocation in Early Modern English literature

The use of word collocation in the analysis of texts has many potential applications, such as aiding in the identification of authorship of anonymous texts, or establishing the semantic categories which are psychologically associated with one another.

Applying the HTST to this work aids and speeds the process of identification of these collocates because of the thorough and fine-grained analysis which the tagger is capable of achieving. The tagger builds on its predecessor, the highly successful USAS tagger, to allow the automatic disambiguation of word meanings (Rayson et al 2004a, Rayson 2008). This disambiguation therefore allows researchers to focus on the 'correct' meanings of a form for their research, rather than have to sift their data in order to remove homonyms which are irrelevant to their work. The current investigation is also of importance because it allows searches to be performed for semantic categories in different levels of granularity. The levels of aggregation allowed by the *Historical Thesaurus* hierarchy and the newly created thematic category set allow the user to specify different degrees of specificity in the results.

3 The corpus

The EEBO corpus transcribed section (EEBO-TCP) as available at April 2014 has been semantically tagged using the HTST as part of the SAMUELS project[2]. The methodology used for the tagging incorporates the normalization of spelling variants using VARD 2.6[3] (Baron and Rayson 2008), which improves the accuracy of the tagging to an extent not previously achievable. The corpus in its tagged state will be made available through a new version of the Wmatrix tool[4] and the Brigham Young website[5], allowing other researchers to conduct work on it without having to go through the process of tagging it themselves.

The EEBO corpus contains a copy of almost every extant work printed in English between the years 1473 and 1700. As such it is a vast and important resource for scholarship on literature and language in the late medieval and early modern period, which is yet to be fully exploited, especially with the kinds of automatic analysis which are being trialled through the SAMUELS project. It is hoped that this will be a test case for the types of powerful analysis which can be achieved when advanced digital humanities software is applied to such an extensive and important corpus.

Through the use of the HTST it can be recognized that words in particular semantic categories collocate with the ideas of *theft* and *authority*. For the former, particular clusters of words are to be found which are to do with animals (such as rat, vulture, worm), nationalities (such as Tartar, Hungarian), and violent action (such as wring, torture, wrack). For *authority*, frequent collocates are words related to physical position (such as elevate, higher, upper), strength (such as mighty, strong), and possession (such as wield, hold).

[2] http://www.gla.ac.uk/samuels/
[3] http://ucrel.lancs.ac.uk/vard/
[4] http://ucrel.lancs.ac.uk/wmatrix/
[5] http://corpus.byu.edu/

4 Semantic annotation

Semantic tagging and annotation is, we argue, the best solution we have to address the problem of searching and aggregating large collections of textual data: at present, historians, literary scholars and other researchers must search texts and summarize their contents based on word forms. These forms are highly problematic, given that most of them in English refer to multiple senses – for example, the word form "strike" has 181 *Historical Thesaurus* meaning entries in English, effectively inhibiting any large-scale automated research into the language of industrial action; "show" has 99 meanings, prohibiting effective searches on, say, theatrical metaphors or those of emotional displays. In such cases, much time and effort is expended in manually disambiguating and filtering search results and word statistics.

To resolve this problem, we use in this paper an intermediate version of the Historical Thesaurus Semantic Tagger, which is in development between the Universities of Glasgow and Lancaster. HTST is a tool for annotating large corpora with meaning codes from the *Historical Thesaurus*, enabling us to search and aggregate data using the 236,000 precise meaning codes in that dataset, rather than imprecise word forms. These *Thesaurus* category codes are over one thousand times more precise than USAS, the current leader in semantic annotation in English corpus linguistics. [6] The system automatically disambiguates these word meanings using existing computational disambiguation techniques alongside new context-dependent methods enabled by the *Historical Thesaurus*' dating codes and its fine-grained hierarchical structure. With our data showing that 60% of word forms in English refer to more than one meaning, and with some word forms referring to close to two hundred meanings, effective disambiguation is essential to HTST.

5 Methodology

The EEBO corpus was lemmatized and then processed through the HTST annotation system, resulting in texts with each word being annotated with a Historical Thesaurus meaning code. We then used the Wmatrix4 user interface to search for semantic categories which were of interest, in this case *Theft* and *Authority*. The results were then examined to produce a listing of the categories which regularly produced collocates with words from the previously selected categories. Our comparison was based on a log-likelihood significance measure first brought to the attention of corpus linguists by Dunning (1993) as a collocation measure, which identifies, to an acceptable degree, those semantic domains which are mentioned unusually frequently in our texts by comparison to the corpus, and therefore indicates a text's "key" domains (where the log-likelihood values are greater than around 20; Rayson et al. 2004b) In addition, we have taken into account range, dispersion and effect size measures.

In order to aid analysis, a set of thematic categories has been created to accompany the *Historical Thesaurus* categories. This thematic categorization offers a significantly reduced set of headings for which a researcher may wish to search. These headings are at a more 'human scale' than that of the *Thesaurus* as a whole, because it focuses on providing headings which are not above or below the level of detail (e.g. 'Biology' on the one hand, and 'Proto-organism as ultimate unit of living matter' on the other) which would be most relevant to the way in which a typical human categorizes the world around them on a day-to-day basis.

The resulting collocates were thus grouped according firstly to the thematic category set. The resulting list showed that words which collocated with the categories of 'Taking surreptitiously/Theft' (AW16) were to be found in animal categories (e.g. 'Birds' AE13 and 'Order Rodentia (rodents)' AE14h) 'Nations' (AD15), and 'Food' (AG01). Additionally, for 'Authority' (BB) the most commonly collocated categories were 'Strength' (AJ04e), 'Position' (AL04a), and 'Possession/ownership' (AW01a).

6 Conclusion

The results produced offer insight into the ways in which the subjects of *theft* and *authority* were conceptualized in the Early Modern period. There is a clear association of theft with animals and with foreign nationals who are often viewed in a pejorative light. On the opposite side of the spectrum, the authorities which try to control the behaviour of subjects are linked with the ideas of strength and possession.

In many of these cases, more close reading as well as reference to the *Historical Thesaurus* and the *Oxford English Dictionary* indicates that many of these collocations are the result of a metaphorical link between the two subjects with, for example, thieves often being described as if they are vermin animals. These links may be further instantiated and testing of their strength possible when the *Mapping Metaphor* project [7], also employing *Historical Thesaurus* data and running at the University of Glasgow, is able to release its data.

[6] http://ucrel.lancs.ac.uk/usas/

[7] http://www.gla.ac.uk/metaphor/

7 Acknowledgements

We acknowledge the support of the SAMUELS project funded by the AHRC in conjunction with the ESRC (grant reference AH/L010062/1), see http://www.gla.ac.uk/samuels/.

References

Baron, A. and Rayson, P. 2008. "*VARD2:a tool for dealing with spelling variation in historical corpora"*. In: Postgraduate Conference in Corpus Linguistics, 2008-05-22, Aston University, Birmingham.

Dunning, T. 1993. "Accurate methods for the statistics of surprise and coincidence". Computational Linguistics 19(1). 61–74.

EEBO. See http://eebo.chadwyck.com/home [accessed 13th January 2015]

Kay, C., Roberts, J., Samuels, M., and Wotherspoon, I. (eds.). 2015. *The Historical Thesaurus of English*, version 4.2. Glasgow: University of Glasgow. http://www.gla.ac.uk/thesaurus

Rayson, P. 2008. "From Key Words to Key Semantic Domains". International Journal of Corpus Linguistics 13.4. 519-549.

Rayson, P., Archer, D., Piao, S. L., and McEnery, T. 2004a. "The UCREL semantic analysis system". In Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25th May 2004, Lisbon, Portugal, pp. 7-12.

Rayson, P., Berridge, D., and Francis, B. 2004b. "Extending the Cochran Rule for the Comparison of Word Frequencies between Corpora". 7th International Conference on Statistical Analysis of Textual Data.

Does Corpus Size Matter? "Exploring the potential of a small simplified corpus in improving language learners' writing quality"

Wael Alharbi
Yanbu University College
whmalh@gmail.com

1 Introduction

The purpose of this submission is to provide a detailed description of the experience of 25 university level students with a small-simplified English corpus that we created as a digital source for linguistic help while writing. Different layers of data collection tools were used to evaluate the participants' experience with the corpus. Not only the participants' writing quality has improved as a result of allowing them to consult the corpus, their attitudes along with the quantity and quality of the corpus queries have improved significantly.

2 Background

According to the British Council (2013), more than 1.5 billion students around the world are learning English as a subject. This number, which is expected to increase to two billion over the next few years, is based on institutions of higher education that instruct through the medium of the English language (Graddol, 2006). As Seidlhofer states: "far more people learning English today will be using it in international contexts rather than in just English-speaking ones" (2011: 17). Currently, there are more second-language English speakers than native-English speakers.

It is indisputable that writing is one of the most important language skills. According to Warschauer (2010), it is an essential skill for students in schools and universities, and for continuous professional development. It is also necessary for EFL/ESL learners for a number of reasons. First, writing well is a fundamental skill for academic or professional success, and in non-English-speaking countries it assists employability and facilitates university education. Second, writing can play an important role in developing learners' academic language proficiency as they become more ready to explore advanced lexical or syntactic expressions in their written work.

Although writing is one of the most important language skills, many students around the world consider it to be the most difficult (Trang, 2009). Witte (2007) too, states that students in her study showed little interest in classroom writing activities

and assignments. Many studies have also shown that it can be difficult to motivate language learners when it comes to writing (Kajder and Bull, 2003; Davis, 1997). Furthermore, according to Mat Daud & Abu Kassim (2005) and Yih & Nah (2009), students' writing performance is anxiety-provoking as a result of their lack of writing skills.

3 Why use corpora in L2 writing classes?

The use of corpora is considered particularly useful in the L2 writing class for a number of reasons (Yoon, 2014; Flowerdew, 2010; O'Sullivan & Chambers, 2006; Yoon & Hirvela, 2004). Corpora are mainly collections of written discourses that expose users to the features and patterns of written language. Learners can discover vocabulary, word combinations and grammatical patterns along with frequencies. Learners can consult them at any stage of the writing process to check if their writing is accurate, if it conveys the intended meaning and/or to find alternatives.

However, not all the studies in corpus linguistics agree that the presence of corpora in L2 classroom can be of great help. The participants of some studies (Turnbull & Burston, 1998; Chambers & O'Sullivan, 2004; Sun, 2007) found that corpus consultation was difficult because it was time consuming to sort through concordance examples and identify relevant ones. They also reported that it was frustrating not to understand all concordance examples and to formulate proper search terms. Some of the participants of these studies also reported their dissatisfaction with corpora simply because they were overwhelmed by the number of examples they got when they searched in the corpora. To avoid these disadvantages, especially with beginners and low proficiency learners, some researchers recommended the introduction of small corpora as they are easier to manage (Chambers, 2007; Yoon, 2014).

In an era when most learners are considered as 'digital natives' and where almost every aspect of our life seems to be governed by technology, introducing corpora into L2 classroom could become a necessity rather than luxury. With computers mediating the writing process, learners need look for linguistic help in the computers hoping to find the information they need at few mouse clicks and keyboards strokes.

In this study we wanted to put these recommendations into question and see if compiling a small corpus that is made of simple English texts and then giving low proficiency language learners access to it can generate positive results. To see if giving learners access to small and simplified corpus would have a positive impact on their experience with the corpus. We therefore had two research questions:

RQ1: What are the attitudes of students to towards the corpus? Do they change over the three phases?

RQ2: Does the quantity and success of the queries change over time?

4 Methods

The participants were 25 Saudi students studying at a university that uses English as a medium of instruction in Saudi Arabia. All of them were competent users of computers, but never heard or used corpora before. The corpus we compiled was extracted from The Voice of America (VOA) Learning English website[8]. All the VOA Learning English content is composed of the most frequently used 1500 words of English. We collected different texts of different genres and created a file consisting of 56,942 tokens. This file then became the core for our small and simplified corpus which used *Antconc* as the concordance for the VOA corpus.

For 12 weeks, we introduced the VOA corpus to our participants over three phases with a new training program being introduced during each phase. During each phase, the participants were asked to write about a topic and were encouraged to consult the VOA corpus for their different language problems. The writing process at each phase was captured using a screen recording software. An attitudes survey was distributed after each writing task. In order to answer the research questions and evaluate the participants' experience with the VOA corpus, we used three lenses through which we hoped to see three different layers of data. The three lenses were:

- The participants' attitudes (by the attitudes survey)
- Their real practices (by the screen video recordings)
- The efficiency of the tool (by the screen video recordings)

We adopted a three-phase gradual training and in each phase we trained them differently. The surveys and recordings were then analysed and the results were crosschecked.

5 Results and discussion

We asked the participants to read the statements in the attitudes survey and rate them On a scale from **1** to **5**, where **1** is "***Strongly Agree"*** and **5** is "***Strongly Disagree"***.

As Figure 2 below shows, the participants' gradual agreement with the statements; ***"VOA corpus gave me confidence in my writing"*** and

[8] www.learningenglish.voanews.com

"VOA corpus made the writing task interesting". For both statements, they gave an average rating of 2.68 indicating partial disagreement with the statement. At phase 2 the attitudes seems to improve a little, but still, they don't agree with it. During the third phase, we can see a good improvement in their attitudes with an average rating of 4 for both statements.

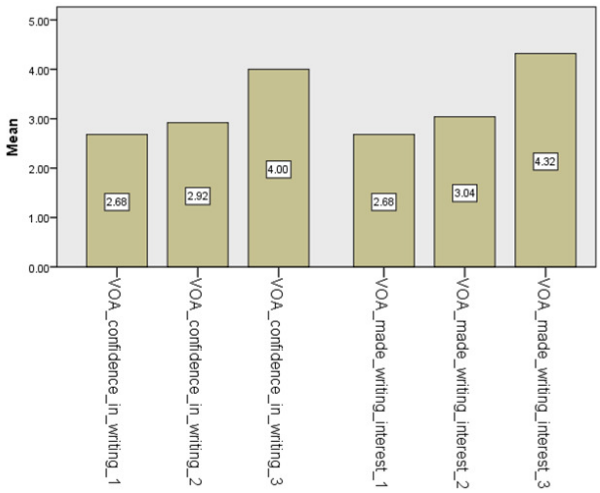
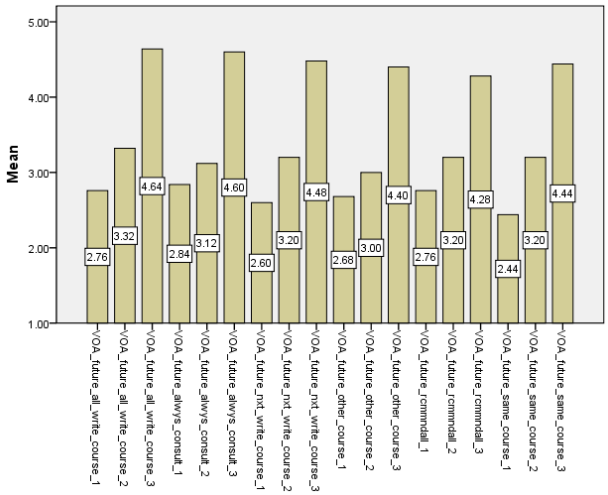


Figure 2: Interest and Confidence in L2 Writing When Using VOA

In the attitudes survey, we also asked them other questions regarding recommending the VOA corpus to others and for future use. Their answers followed a similar trend starting from low to high. However, the difference here was in the third phase where the bars are higher indicating stronger agreement with the statements.

Figure 3: Recommending VOA for Future Use



The survey results above show an increasing trend leading upwards. The results also show that the participants have more positive attitudes as they progressed towards the end of the study with the biggest increase seen in phase three. To answer the first research question, the survey results showed extremely positive attitudes towards the VOA corpus during phase three with a gradual improvement in the attitudes.

The second lens through which we looked at the students' experience was through the screen video recordings. Figure 4 below shows the total number of queries conducted in the VOA corpus during the first phase of the study. We can see three different labels:

- **Success**: when a student conducts a query in the VOA and then uses the search result successfully.
- **Wrong**: when a student conducts a query in the VOA and then uses the search result unsuccessfully.
- **Abandonment**: when a student conducts a query in the VOA and then decides not to use the search result.

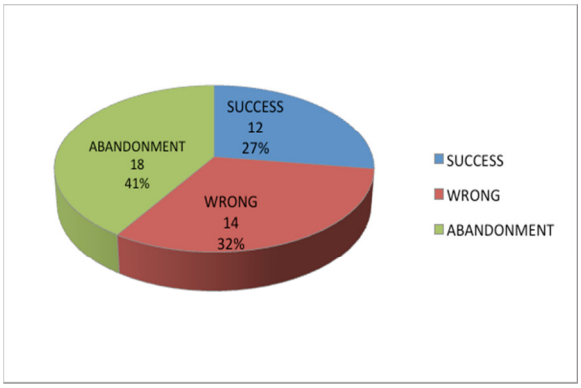


Figure 4: Level of Success of the VOA Corpus during Phase One [44]

As Figure 4 above shows, the participants conducted a total 44 queries during the first phase of the study only 27% of which were successful. The

remaining 73% of the searches were either wrong (32%) or Abandoned (41%).

During the second phase, however, the frequency of queries has increased to 60 searches, but the quality has decreased as only 22% of the queries were successful.

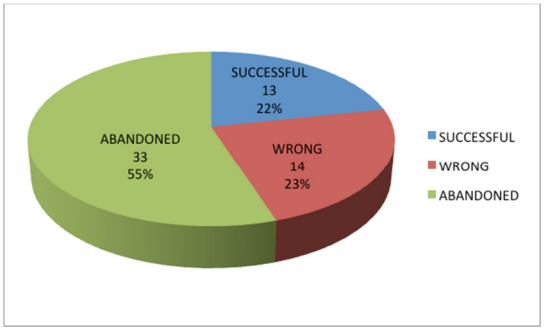


Figure 5: Level of Success of the VOA Corpus during Phase Two [60]

During the third phase of the study, the corpus searching behavior took an opposite trend to the second phase. In the third phase the participants conducted only 14 searches most of which were successful (64%).

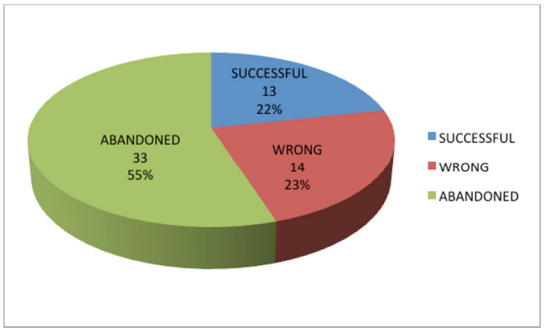


Figure 6: Level of Success of the VOA Corpus during Phase Three [14]

Although the results above show an increase in the percentage of the successful queries during phase three, success decreases during phase two and frequency of searches decrease during phase three. This indicates that the VOA simplified English corpus is a good resource for improving L2 learners' attitudes towards the use of corpora in L2 classroom, but because of the corpus size being very small, the participants couldn't make the most out of it.

6 Conclusion

As a whole, the results of our exploratory study revealed that the VOA simple English corpus with proper training can be as a viable linguistic reference tool for enhancing learners' attitudes and the linguistic aspects of L2 writing. However, no matter how enthusiastic practitioners are about introducing an intervention into classroom, different layers of data is what shows the full picture. In our study, the survey results showed the biggest improvement of the attitudes taking place during phase three, while the analysis of the screen recordings showed a decrease in the number of times the participants consulted the VOA corpus. Depending on one source of data could blur the picture especially when the source of the data come from the participants' self-reports.

In order to investigate this great source of linguistic help, we recommend that researchers may develop the VOA corpus by probably opting for a more user-friendly interface and by increasing the size of the corpus. Evaluating the quality of the written texts may will add another layer of data and may eventually show a clearer picture of the learners' experience with the resources.

References

British Council. 2013. Culture Means Business. Available at http://dera.ioe.ac.uk/18071/14/bis-13-1082-international-education-global-growth-and-prosperity-analytical-narrative_Redacted.pdf

Chambers, A. 2007. Popularising corpus consultation by language learners and teachers. In E. Hidalgo, L. Quereda, & J. Santana (Eds.), *Corpora in the foreign language classroom* (pp. 3–16). Amsterdam, Netherlands: Rodopi.

Chambers, A., & O'Sullivan, I. 2004. Corpus consultation and advanced learners' writing skills in French. *ReCALL*, 16(1), 158–172.

Davies, S. 2003. Content-based instruction in EFL contexts. *The Internet TESL Journal*, 4 (2). Re

Graddol, D. 2006. *English Next: Why Global English May Mean the End of 'English as a Foreign Language'*. London: British Council.

Sun, Y.-C. 2007. Learner perceptions of a concordancing tool for academic writing. *Computer Assisted Language Learning*, 20(4), 323–343.

Turnbull, J., & Burston, J. 1998. Towards independent concordance work for students: lessons from a case study. *On-Call*, 12(2), 10–21.

Yoon, H. 2014. Direct and indirect access to corpora: An exploratory case study comparing students' error correction and learning strategy use in L2 writing. *Language Learning & Technology* 18(1), 96–117. Retrieved from http://llt.msu.edu/issues/february2014/yoonjo.pdf

Yoon, H., & Hirvela, A. 2004. ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing,* 13(4), 257–283.

Seeing corpus data: Lessons from visualising UK press portrayals of migrants

William L Allen
University of Oxford
william.allen@compas.ox.ac.uk

1 Introduction

Increasingly, researchers in the arts and humanities as well as the social sciences are visually communicating the results of their studies to the wider public. Through visualisation, or 'the representation and presentation of data that exploits our visual perception abilities in order to amplify cognition' (Kirk 2012), researchers can highlight important findings or emergent trends. Interactive visualisations also enable users to explore data and analyses to meet their own curiosities and interests.

As text corpora and corpus methods become more available to researchers, there is great potential for conducting and sharing analyses of text through visualisations. Some studies in linguistics have introduced visual networks as both modes of analysis and communication of descriptive results (Di Cristofaro 2013). But there remain important lessons for linguists that go beyond the technical 'how-to'. What issues arise when large corpora are visualised by and for non-academic groups like journalists or public policy think tanks? What should researchers be aware of as they begin representing their textual data? How do users react when they are given an opportunity to 'see corpus data', possibly for the first time?

2 Project background, data, and methods

As part of the AHRC-funded project *Seeing Data: Are Good Big Data Visualisations Possible?*,[9] the research team of Helen Kennedy (Principal Investigator, University of Sheffield), William L Allen (Co-Investigator), Andy Kirk (Consultant Researcher, Visualising Data Ltd), and Rosemary Lucy Hill (University of Leeds) examined what makes visualisations particularly effective from the perspectives of their producers (e.g., designers) and a wide range of users.

One aspect of the project involved visualising a corpus containing UK newspaper coverage of immigration, built by The Migration Observatory at the University of Oxford. This corpus included, as far as possible, articles from all national British newspapers which mentioned any of a number of immigration-related terms from 2006 to 2013.[10] This corpus was stored in the Sketch Engine, a piece of web-based lexicographical software that tags words with their parts of speech and allows users to perform a range of corpus linguistic analyses on corpora, including collocational analysis. The corpus was divided into annual subcorpora as well as by publication type: tabloids, midmarkets, and broadsheets. Since the Observatory informs public debate about international migration using data and evidence that is transparently communicated, *Seeing Data* aimed to help the Observatory understand whether visualisation of its data could also help achieve this purpose.

From June-August 2014, the project team enlisted Clever Franke, a professional design firm based in Utrecht, to build two bespoke visualisations for the Observatory, one of which was based on the textual dataset. Then, from August-November 2014, the team conducted nine focus groups (one of which was a pilot) involving 46 participants in Dumfries and Galloway, Lincolnshire, Leeds/Bradford, and Oxfordshire. These places—a mix of rural as well as urban places—were intentionally selected because the team felt they represented regions with different migration backgrounds and might draw participants with different reactions to visualisations about migration. Some cities had more recent experiences of migration (e.g., Boston in Lincolnshire) while others had multiple generations of immigration (e.g., Leeds and Bradford).

During the focus groups, participants viewed up to nine visualisations featuring a range of topics and styles that the team had selected. Then they were asked to share their thoughts about what they had felt or learned. Two of the visualisations were the designs by Clever Franke, although participants were not made aware of Co-I Allen's affiliation with the Observatory prior to the focus groups.

3 Generating visualisations from corpus data: experiences of The Migration Observatory

The Observatory's ongoing experience of transforming corpus linguistic insights into visual representations, including the work completed in *Seeing Data*, highlighted several important issues to which researchers considering visualisation should be attuned. First, understanding the motivation for a visualisation in the first place is vital because it informs future choices about design, content, and interactivity. For example, in an earlier pilot project that investigated media portrayals of migrants from

[9] The project was funded under the AHRC's Digital Transformations Call from January 2014 to March 2015. For more information about the project, visit www.seeingdata.org .

[10] The search string and general approach was based on Gabrielatos and Baker (2008).

2010 to 2012, the Observatory published an interactive visualisation using Tableau Public as seen in Figure 1 (Allen and Blinder 2013). It allowed users to customise the visual output of collocates along several dimensions including newspaper type and migrant group. The rationale for including this kind of feature stemmed from the Observatory's value of transparency: users could see and customise the collocation results for themselves.

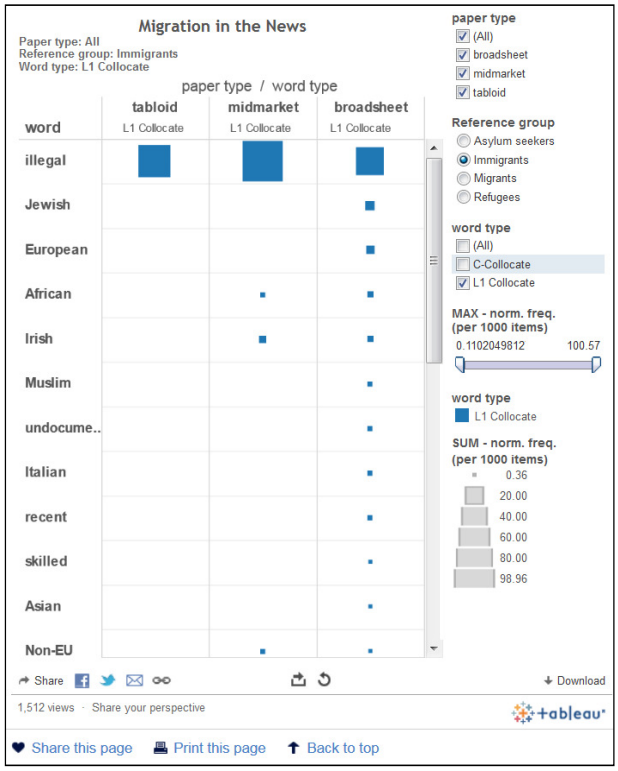


Figure 1. Screenshot of *Migration in the News* Visualisation

Understanding why organisations or researchers would want to visualise corpora also reveals the extent to which they are located in professional, normative, or political contexts. For what purposes will these visualisations be used? Are they designed to enable users to access or 'read' discrete values, or are they trying to evoke certain feelings or emotions (Kirk 2014)? Does the designer have a particular 'style', sense of mission, or practical way of working which may influence the outcome?

Second, after considering the rationale for a corpus visualisation, there is the issue of choosing which linguistic features to visualise—and how. In the case of the media corpus, frequencies of the key terms 'immigrants', 'migrants', 'asylum seekers', and refugees' over time and by publication type were plotted as a line graph. This was accompanied by points showing key moments in UK migration policy change or British politics like elections. The combination of these features was intended to provide additional context to users as they explored the data. Similarly, the corpus analysis also revealed which adjectives most regularly collocated with each migrant group. This was visually represented by showing the top 100 collocates of each term, with the relative strength of a collocation indicated by a more saturated colour. These collocates were divided by publication type as well.

The experience of visually representing linguistic features also revealed a third issue: communication among people using different professional 'languages'. For example, working with designers who had little to no background in linguistics raised crucial questions around the meaning of collocation and the provenance of the data. Mutual understanding of the project rationale and key modes of analysis to be visualised was vital.

4 Interacting with a visualisation: user intentions and (dis)trust

But the perspectives of designers and commissioning organisations are only part of the story. The other parts involve users who interact with the visualisation. During the course of the focus groups, two issues emerged which have implications for the ways that corpora are effectively visualised.

The first centred around user intentions, or the expectations that users had for the visualisation. What motivation did they have to interact with this visualisation? In some cases, it was professional or research interest: several participants were working in media or communications, for example. Others had personal experience of being a migrant in the UK and connected with the subject matter. These kinds of intentions and potential audiences are important for linguists to consider as they visualise their analysis because they can impact how the subsequent presentation is interpreted.

Secondly, participants raised the issue of (dis)trust, especially with this particular visualisation of media texts. It was apparent that the political importance of immigration, particularly in some of the regions that had experienced recent migration, impacted the reception of the visualisation and the underlying corpus methods. For example, despite the presence of explanatory text about the methods used, as well as their comprehensiveness and limitations (Allen 2014), some participants expressed scepticism over the intentions of the visualisation: the fact it was a corpus of media outputs suggested to some that it was automatically politically biased and motivated to 'counter' negative portrayals. Yet this was not universally felt: others pointed out that features like the breadth of data and academic branding communicated a sense of trust.

These issues of user intentions and trust exemplify how reception of visualisations is affected

by a number of factors—some of which lie within the control of either visualiser or linguist.

5 Conclusion

Among several objectives, *Seeing Data* aimed to provide the Observatory with greater insight into how visualisation of large datasets like corpora about migration news coverage can be achieved. It provided at least five important lessons: (1) consider the aims and rationale of a visualisation in the first place, before decisions about design are made; (2) link choices of linguistic features to visualise with design options; (3) build time into the work for developing clear communication among academics and visualisers who may come from different backgrounds; (4) consider what the intended audiences of the visualisation will gain; and (5) acknowledge how the topics of corpora and their visual presentation can lead to judgments of whether to believe the visualisation at all. These lessons are applicable not only to linguists interested in visualising their work, but also to broader humanities and social science researchers working with text as a form of data.

Acknowledgements

The author would like to acknowledge the collective contributions of the *Seeing Data* team and advisory board in the development, fieldwork, and analysis stages.

References

Allen, W. (2014). *Does Comprehensiveness Matter? Reflections on Analysing and Visualising Uk Press Portrayals of Migrant Groups*. Paper presented at the Computation + Journalism Symposium, Columbia University, New York City.

Allen, W., & Blinder, S. (2013). Migration in the News: Portrayals of Immigrants, Migrants, Asylum Seekers and Refugees in National British Newspapers, 2010 to 2012 *Migration Observatory Report*. University of Oxford: COMPAS.

Di Cristofaro, M. (2013). *Visualizing Chunking and Collocational Networks: A Graphical Visualization of Words' Networks*. Paper presented at the Corpus Linguistics Conference 2013, Lancaster University.

Gabrielatos, C., & Baker, P. (2008). Fleeing, Sneaking, Flooding a Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the Uk Press, 1996-2005. *Journal of English Linguistics,* **36**(1), 5-38.

Kirk, A. (2012). *Data Visualization: A Successful Design Process*: Packt Publishing Ltd.

Kirk, A. (2014). Recognising the Intent of a Visualisation. http://seeingdata.org/getting-gist-just-enough/

A review of morphosyntactic analysers and tag-sets for Arabic corpus linguistics

Abdulrahman AlOsaimy
University of Leeds
scama@leeds.ac.uk

Eric Atwell
University of Leeds
e.s.atwell@leeds.ac.uk

1 Introduction

Geoffrey Leech applied his expertise in English grammar to development of Part-of-Speech tagsets and taggers for English corpora, including LOB and BNC tagsets and tagged corpora. He also developed EAGLES standards for morphosyntactic tag-sets and taggers for European languages. We have extended this line of research to Arabic: we present a review of morphosyntactic analysers and tag-sets for Arabic corpus linguistics.

The field of Arabic NLP has received a lot of contributions in the last decades. Many analysers handle its morphological-rich problem in Modern Standard Arabic text, and at least there are six freely available morphological analyzers at the time of writing this paper. However, the choice between these tools is challenging. In this extended abstract, we will discuss the outputs of these different tools. We show the challenge of comparing between them.

The goal of this abstract is not to evaluate these tools but to show the differences. We aim also to ease the building of an infrastructure that can evaluate every tool based on common criteria and produce a universal pos-tagging.

2 Presentation of morphological analysers

- **BAMA:** A widely-known Perl-based freely available Arabic morphological analyser by Tim Buckwalter. The analyser used in this research is version 1.3. Later versions needs an LDC licence and therefore not considered in this comparison.

Outputs: POS tag, gloss, voweled word and stem. The tagset of Buckwalter is about 70 basic subtags, and they can be combined to form more complex tag such as: IV_PASS which means imperfective passive verb. Those tags include features of verbs like person, voice, mood, aspect and its subject like gender and number. It also includes features of nominal like gender, number, case and state. BAMA provides a list of different analysis with no disambiguation of them.

- **Mada:** a freely available toolkit that tokenizes, pos-tags, lemmatize, stems a raw Arabic input. This toolkit, its successor MADAMIRA disambiguates the analyses by showing the probability of each analysis.

Outputs: POS tag, gloss, voweled word, stem and the word lemma. The output tagset can be one of four different POS tagsets: ALMORGEANA, CATiB, POS:PENN, Penn ATB, or Buckwalter. Features of verbs like person, voice, mood, aspect and its subject like gender and number are explicitly provided. Same for features of nominal like gender, number, case and state. MADA provides a list of different analysis each with a probability. The higher is the more likely one.

- **MadaAmira**: is the Java-based successor of Mada that combines Mada and Amira tools. It adds some aspects from Amira tool.

Outputs: In addition to the output of Mada, the base phrase chunks and named entities can be provided.

- **AlKhalil**: **"**a morphosytactic parser" of MSA that is a combination of rule-based and table-lookup approach.

Outputs: AlKhalil is different as it all output is a table-like provided in Arabic sentence that describe the morphological analysis of each word. The table have POS-tags, prefix, suffix, pattern, stem, root and voweled word columns. Features of verbs like voice, transitivity and aspect are extractable. However the mood and person is not explicitly provided neither its subject if it a suffix. Nominal features are also extractable. In addition AlKhalil provides the nature of the noun, word root, and verb form.

- **Elixir:** is a morphology analysers and generator that reuse and extends the functional morphology library for Haskell.

Outputs: Elixir uses a custom output format including gloss, voweled word, root, stem, pattern, and a 10-letters word that describes the POS tag and all words features as Mada.

- **AraComLex**: is an open-source finite-state morphological processing toolkit.

Outputs: AraComLex provides the main POS tags categories: prep, conj, noun, verb, rel, adj … etc. For nominals, it provides its classification class (13 classes), number, gender, case, and whether it is human or not. For verbs, it provides number, gender, person, aspect, mood, voice, transitivity and whether allows passive or imperative.

- **ATKS**: is web-based service of NLP components targeting Arabic language that includes "full-fledged" morphological analyser (Sarf) and part-of-speech (POS) tagger.

Outputs: Like Buckwalter tagset, ATKS provides complex tags that encompass nominal and verb features. All features are extractable from pos-tags. Sarf provides a list of features like: stem, root, pattern, discretized token, isNunatable and probability of each analysis.

- **Stanford NLP tools:** open-source software in Java that has a segmenter, pos-tagger and parser of Arabic text.

Outputs: The output of Stanford parser and pos-tagger is Bies tagset which is used for Arabic Penn Treebank. This tagset is linguistically coarse (Habash 2010) and therefore many features are missing. The features that are extractable are aspect (unless it is passive as perfect and imperfect verbs share the same tag), number (singular or plural only) and voice.

- **Xerox:** web-based morphological analyser and generator built using Xerox Finite-State Technology.

Outputs: The output of Xerox analyser includes POS tag, English gloss, root, verb form and verb pattern. Features of verbs like person, voice, mood, aspect and its subject like gender and number are provided. Same for features of nominal like gender, number, case and state.

- **QAC**: the Quranic Arabic Corpus is a linguistic resource that includes segmentation and pos-tagging the Quran text. We used this resource as the gold standard for evaluating other tools as it has been verified by experts in Arabic language.

3 Work

We built an infrastructure for parsing all results from the tools mentioned above. For every analysis of a word, we parsed the tags associated with it and extracted the features (if possible) of the nominals and verbs (Fig 1).

We plan to benchmark every tool by comparing its results to the Quranic Arabic Corpus. For every feature that the QAC provides, we will find the accuracy, precision, and recall of each tool. However, benchmarking needs to first map all part of speech tags to one universal tag set. Another problem is that some tools provide different *unordered* analyses. We plan to find the best analysis that matches the QAC and report the results of that analysis.

Feature	Possible Values	Applied to
Gender	Male/Female	Nomonials & Subj. of verb
Number	Sing./Dual/Plural	Nomonials & Subj. of verb
Case	nominative, accusative, genitive	Nominals
state	Definite or Not	Nominals
Person	First, Second, Third	Verbs
voice	active, passive	Verbs
aspect	perfective, imperative, imperfective	Verbs
mood	indicative, subjunctive, jussive, energetic	imperfective verbs

Table 1 inflectional features in Arabic

4 Challenges

Problem 1: The diverse in the format of the output: Every tool has its own format of output. Alkhalil return a table-like CSV file. Mada and MadaAmira return a text of *feature:value* pairs. However, some tools have more complex output like BAMA that needs to build a custom parser designed specifically for that tool. Therefore, for each tool, we need to translate the custom outputs to an open standard format: JSON. As a consequence, the infrastructure needs to be updated every time one of the tools changes its output scheme.

Problem 2: The availability of some tools: While many researchers published papers about their morphology tools, many of these are either not available or require a licence. For example, although Mada toolkit is freely available, it requires a lexicon tables that are only available with membership of LDC. In addition, some web services such as Xerox are limited to some quotas.

Problem 3: Different segmentation of words: For a valid comparison, words need to be similarly segmented. However, some tools cannot accept segmented text and instead it segments the input text as a preprocessing step.

Problem 4: Extracting features from POS tags: Although some tools do not explicitly present some important features such as gender, number and person, these features can be extracted from the POS tag of that word. However, such handling needs very careful understanding of the POS tags and could produce some errors by such manipulation. Every tool has its own set of tagsets. Tagsets sizes vary wildly. Buckwalter tagset for example can hypothetically reach over 330,000 tags (Habash 2010), while Stanford tagger used Bies tagset that has around 20+ tags. Those tagsets needs to be mapped to one universal tagset in order to be able to compare between them. Mapping will result in many features unknown, or have multiple possible values. In addition, the values of some features do not cover all possible values; number feature in Stanford can be only singular or plural, but in Arabic it could be dual.

Problem 5: Different possible configurations: Mada has different configurations of preprocessing the input text. Different configurations lead to different tokenization, and therefore different analyzing. We chose the default settings, and we will leave comparing different configurations for future work.

Problem 6: Expectancy of input: While some tools expect unvoweled text data (AraComLex), some accept fully or partially voweled such as AlKhalil. ATKS used these short vowels to filter the best analyses if it fits or the diacritics will be ignored. Mada expects the input text to be text-only one sentence per line with no tags or meta data. AraComLex expects every word to be in a single line. Stanford parser expects tokenized words except the definitive AL.

Problem 7: Different Transliteration Schemes: Different tools encode the results in either ASCII or UTF-8. Some use a one-to-one transliteration scheme like Buckwalter transliteration. However, B.W. transliteration received several extensions, and determining which extension can be difficult when tool has a lack or poor user manual. Other tools like Elixir uses ArabTex encoding whose mapping can be two-to-one or has some alternatives.

References

Aliwy, Ahmed Hussein. *Arabic Morphosyntactic Raw Text Part of Speech Tagging System*. Diss. Repozytorium Uniwersytetu Warszawskiego, 2013.

Habash, Nizar Y. "Introduction to Arabic natural language processing."*Synthesis Lectures on Human Language Technologies* 3.1 (2010): 1-187.

Smrž, Otakar. *Functional Arabic Morphology. Formal System and Implementation*. Diss. Ph. D. thesis, Charles University in Prague, Prague, Czech Republic, 2007.

Boudlal, Abderrahim, et al. "Alkhalil Morpho SYS1: A Morphosyntactic Analysis System for Arabic Texts." *International Arab Conference on Information Technology*. 2010.

Dukes, Kais, and Nizar Habash. "Morphological Annotation of Quranic Arabic."*LREC*. 2010.

Atwell, E. S. "Development of tag sets for part-of-speech tagging." (2008): 501-526.

Jaafar, Younes, and Karim Bouzoubaa. "Benchmark of Arabic morphological analyzers challenges and solutions." *Intelligent Systems: Theories and Applications (SITA-14), 2014 9th International Conference on*. IEEE, 2014.

Pasha, Arfath, et al. "Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of

arabic." *In Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland*. 2014.

Habash, Nizar, Owen Rambow, and Ryan Roth. "Mada+ tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization." *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt*. 2009.

Green, Spence, and Christopher D. Manning. "Better Arabic parsing: Baselines, evaluations, and analysis." *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 2010.

Attia, Mohammed, et al. "A lexical database for modern standard Arabic interoperable with a finite state morphological transducer." *Systems and Frameworks for Computational Morphology*. Springer Berlin Heidelberg, 2011. 98-118.

Buckwalter, Tim. "Buckwalter {Arabic} Morphological Analyzer Version 1.0." (2002).

Introductions in Engineering Lectures

Siân Alsop
Coventry University
alsops@uni.coventry.ac.uk

Hilary Nesi
Coventry University
hilary.nesi@coventry.ac.uk

There have been many move analyses of research article introductions, including several that have drawn on academic corpora, starting with (and heavily influenced by) the work of Swales (1981, 1990, 2004). The conventions for structuring research articles are relatively stable and are well-understood by members of the relevant research communities; research article introductions are almost always demarcated by section headings, for example, and typically consist of a series of moves aiming to create a 'research space' for the article to occupy. In many other academic genres, however, there is greater variation in the purpose and structure of introductory sections. Even if they all function to 'introduce the academic work' (Bhatia 1997: 182), practitioners do not necessarily agree about their generic features. Nesi and Gardner (2012: 98) found considerable variation in the role of introductions in student essays, for example, and Bhatia's informants disagreed about the distinctions between *introductions, prefaces* and *forewords* to academic books (1997: 183).

The structural conventions of spoken academic genres are particularly difficult to identify, because moves are not usually labeled in a manner analogous to titles or section headings, and because speech events unfolding in real time are of necessity more disorganized and idiosyncratic than texts carefully drafted for publication or coursework submission. It may be that body language and other visual clues (Yaakob 2013; Yeo and Ting 2014) or 'phonological paragraphs' marked by changes in pitch and intonation (Thompson 2003) signal transitions between stages in lectures, but only small samples of lectures have been analysed with this in mind because the major spoken academic corpora are not annotated for visual or prosodic features. Instead most analyses of lecture openings have been undertaken using models similar to those used for written academic genres. Building on Thompson's Lecture Introduction Framework (1994), they treat introductions as a subgenre of the academic lecture, and identify two or three main introductory stages involving warming up (housekeeping and previewing), setting up (in terms of topic, scope and aims) and putting the topic into context (in terms of its importance, and the students' prior knowledge)

(Lee 2009; Yaakob 2013; Shamsudin and Ebrahimi 2013). The assumption seems to be that introductions perform noticeably different types of discourse function to those in the main explanatory body of the lecture, and that they are the "preliminary part before the lecturer embarks on a new topic or subtopic for the lecture proper" (Yeo and Ting 2014). Indicators of the transition from introduction to the explanatory body of the lecture might be the presence of a lengthy pause followed by a boundary marker such as *right* or *okay* (Thompson 1994; Lee 2009), or the first presentation of new information in the lecture (Shamsudin and Ebrahimi 2013; Yeo and Ting 2014).

The 'beginning, middle and end' model into which lecture introductions are fitted in these studies does not accord with Young's findings from phasal analysis, however (1994). Young argues instead that there is no specifically introductory phase, and that preview, conclusion and evaluation phases are interspersed with theory, example and interaction phases discontinuously, throughout the lecture.

We argue that the notion of 'introduction' is indeed not a very satisfactory one when applied to academic lectures. Our analysis draws on data from 76 lectures from the Engineering Lecture Corpus (ELC, www.coventry.ac.uk/elc), which has been annotated for the pragmatic functions of housekeeping (timetabling and examination notifications, assignment collection and return etc.), summarising (previewing and reviewing lecture content), humour and storytelling. On qualitative inspection we were unable to find a reliable way of identifying a shift between introductory material and the main body of the lecture, so rather than selecting text up to a transition point marked by a pause, or by lexical or phonological features, we examined the distribution of various linguistic and pragmatic features within the first 10% of tokens from each lecture in the corpus. Data was extracted based on a percentage (rather than raw) token count to accommodate small variations in lecture length.

Within these opening sections, we found that lecturers often returned to a preview summary following the delivery of new content. This meant that signals of new information did not reliably indicate a shift from introductory material to main body content. In one lecture on electrical theory, for example, the first ten percent (536 tokens) consisted of a series of summaries interspersed with other discourse functions *and* new information, in the following pattern:

> *housekeeping* → *summary* (preview current lecture; review previous lecture; preview future lecture) → *housekeeping* → *summary* (preview current lecture) → *housekeeping* → *humour* (irony/sarcasm) → *story* (narrative) → *summary* (preview current lecture) → *humour* (irony/sarcasm) → *summary* (preview future lecture) → *new information* → *housekeeping* → *new information* → *housekeeping* → *new information* → *summary* (preview current lecture) → *new information* … (ELC_1021)

Similarly, in this lecture and in others the presence of discourse markers was not a good indication of transition from introduction to the explanatory body of the lecture. Markers such as *right* and *okay*, followed by pauses or hesitation devices, did mark transitions from one function to another, but did not necessarily precede the presentation of new information.

As may be expected, a large amount of all housekeeping occurs in the first 10% of most of the ELC lectures, and there is usually some type of summarizing in the first 10%, largely consisting of reviewing previous and previewing current lecture content. The distribution of these summaries, however, is far from limited to lecture openings.

Figure 1 gives a sense of where and for how long summaries and housekeeping (shown in light grey) occur in comparison to humour and storytelling (in dark grey) and the fundamental teaching functions of defining, explaining etc. (white space).

Figure 2 shows the distribution of housekeeping phases, and Figure 3 shows the distribution of summaries. The 76 ELC lectures are represented on the y-axis as a stacked bar chart, with 0-100% of the normalized duration (in tokens) shown on the x-axis.

From these visual overviews it is clear that a number of stretches of text which do not perform a preview/review or a housekeeping function are also present in the opening 10% of ELC lectures. Storytelling is not uncommon, and one lecturer spends most of this opening part delivering three consecutive jokes (ELC_1030). Another lecturer simply dives straight in with new information without reference to subsequent lecture content (ELC_2008), and the only signposting offered by one lecturer is "we'll begin with lecture number fifteen on page seventy seven" (ELC_3012).

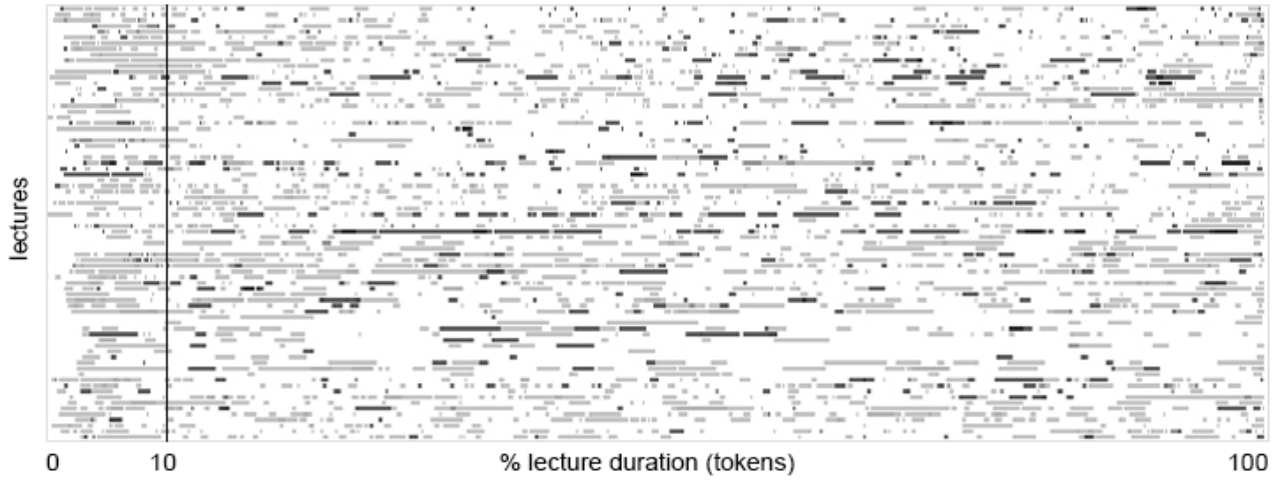


Figure 1: A visualisation of the occurrence and duration of selected discourse functions in the ELC

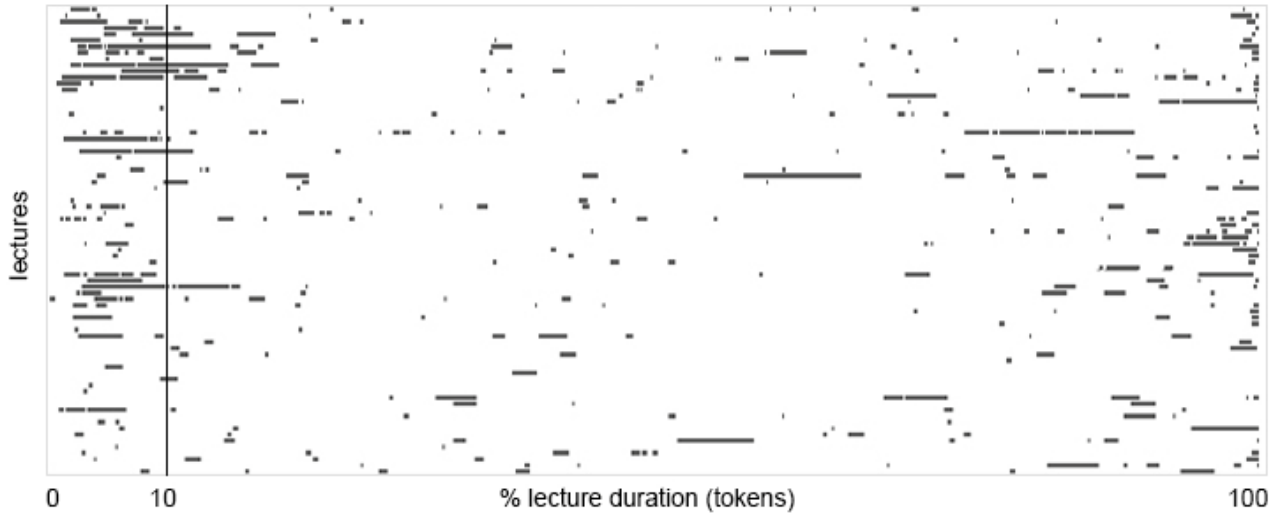


Figure 2: A visualisation of the occurrence and duration of housekeeping in the ELC

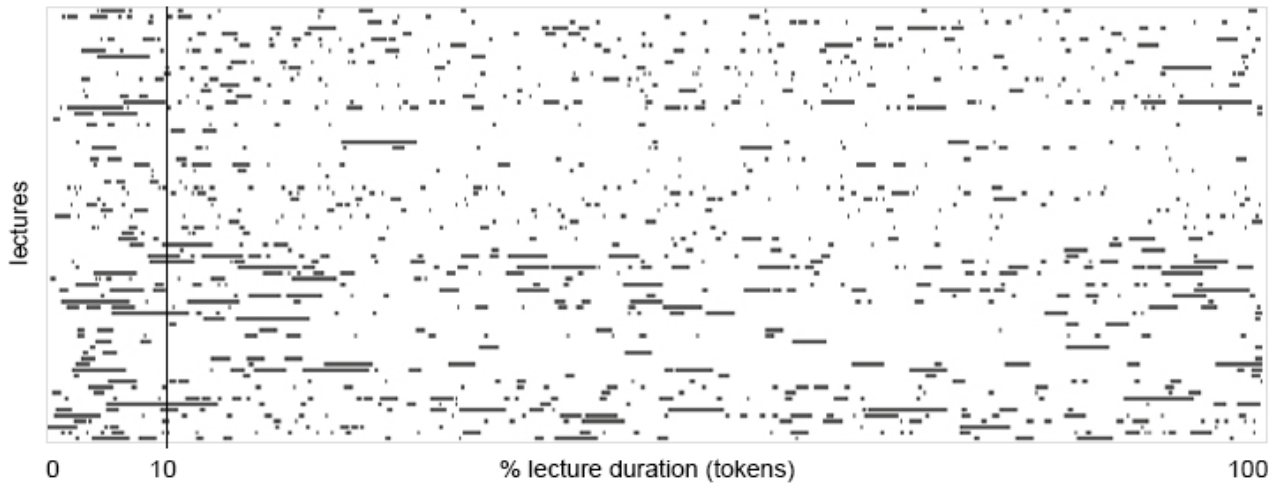


Figure 3: A visualisation of the occurrence and duration of summaries in the ELC

Based on a statistical overview and investigation of examples of common discourse functions that occur in the ELC lecture openings, we argue that in practice the concept of an 'introduction' is not a very useful one for lectures. In most cases: a) introductions are not explicitly marked as such, and although there is signposting this can also recur at later stages in the lecture, and b) the first part of a lecture may realize a number of different pragmatic functions, but these functions also recur at later stages.

What our results flag up is that in terms of EAP

teaching (when giving note-taking instruction, for example) introductions in lectures do not perform the same crucial information delivery function as introductions in written academic texts. Thus the premise that lecture introductions are particularly important for lecture comprehension may well be false – or at least overemphasized.

References

Bhatia, V. K. 1997. "Genre-Mixing in Academic Introductions." *English for Specific Purposes* 16 (3): 181-195.

Lee, J. J. 2009. "Size matters: An exploratory comparison of small- and large-class university lecture introductions." *English for Specific Purposes* 28 (1): 42-57.

Nesi, H. and Gardner, S. 2012. *Genres Across the Disciplines: Student Writing in Higher Education*. Cambridge: Cambridge University Press.

Shamsudin, S. and Ebrahimi, S J. 2013. "Analysis of the moves of engineering lecture introductions." *Procedia - Social and Behavioral Sciences* 70 (0): 1303-1311.

Swales, J. M. 2004. *Research Genres: Explorations and Applications.* Cambridge: Cambridge University Press.

Swales, J. M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Swales, J. M. 1981. *Aspects of Article Introductions*. Language Studies Unit at the University of Aston in Birmingham (reprinted 2011, Michigan University Press).

Thompson, S. E. 2003. "Text-structuring metadiscourse, intonation and the signalling of organisation in academic lectures." *Journal of English for Academic Purposes* 2 (1) 5–20.

Thompson, S. E. 1994. "Frameworks and contexts: A genre-based approach to analysing lecture introductions." *English for Specific Purposes* 13 (2) 171-186.

Yaakob, S. 2013. A Genre Analysis and Corpus Based Study of University Lecture Introductions. Unpublished PhD thesis, University of Birmingham.

Yeo, J-Y. and S-H. Ting. 2014. "Personal pronouns for student engagement in arts and science lecture introductions." *English for Specific Purposes* 34: 26-37.

Young, L. 1994. "University lectures – macro-structure and micro-features". In J. Flowerdew (ed.) *Academic Listening*. Cambridge: Cambridge University Press.

Muslim and Christian attitudes towards each other in southwest Nigeria: using corpus tools to explore language use in ethnographic surveys

Clyde Ancarno
King's College London
clyde.ancarno@kcl.ac.uk

Insa Nolte
University of Birmingham
m.i.nolte@bham.ac.uk

Corpus linguistic tools are used by an increasingly diverse range of academics with no previous expertise in linguistics to support the analysis of their language data. We are particularly interested in the ways in which corpus linguistic tools can aid anthropologists. Our corpus-assisted discourse analytic investigation into religious tolerance therefore uses language-based anthropological data. This data emanates from an ethnographic survey (carried out in southwest Nigeria in 2012-13 and involving more than 2800 participants) gathered as part of the larger research project: 'Knowing each other: everyday religious encounters, social identities and tolerance in southwest Nigeria' (henceforth 'KEO'). This project examines the coexistence of Islam, Christianity and traditional practice in the Yoruba-speaking parts of Nigeria. It focusses on how people encounter other religions and how their identity is shaped by this encounter. It also posits that exploring the beliefs and attitudes of participants about these encounters with the religious 'others' can contribute to achieving these aims.

Our corpus differs considerably from what usually gets referred to as 'corpora' in corpus linguistics for it consists of the answers to all 60 open-ended questions in the above-mentioned survey. These questions asked respondents either to discuss their own or family members' experiences of inter-religious encounter, or their views on hypothetical scenarios relating to inter-religious encounter. In the part of the questionnaire focusing on their own experiences, respondents were asked to explain (if applicable) why they or family members had changed their religion, any attempts others had made to convert them, and the reasons they felt people in general converted. They were also asked for their views on inter-religious marriage and their experiences and views on religious differences between children and parents. Another part of the survey asked them about their experiences of participating in practices associated with other religions or Yoruba traditions, including hypothetical scenarios such as whether they would

visit an alfa (Islamic cleric) or pastor if they had a problem, whether they would allow family members of a different religion to attend a family celebration, and about how they would accommodate family members and friends of different religions at social events. Finally, they were asked what they liked or respected about Islam, Christianity and Yoruba customs, how they would advise religious leaders to behave towards one another, their experiences of religious conflict and how they would suggest such conflict could be prevented. Our corpus is therefore thematic, for it captures discourse about interreligious encounters in Yorubaland in South West Nigeria and we use it to 'investigate cultural attitudes expressed through language' (Hunston 2002: 13-14). Insofar as the reasons for the survey were clear and the data collection was rigorous, we argue that the KEO corpus is homogeneous. The overall corpus is 'bilingual' in that some participants answered in English and others in Yoruba. However, we use the English component of the KEO corpus, i.e. all the answers in English (approximately 300,000 words). The KEO corpus evidently falls under the category of 'ad hoc' specialised corpora (specialised corpora are particularly relevant for a range of social scientists for whom a corpus of general English is not relevant) and is relatively small owing to the large corpora other corpus-driven research, for example, utilises.

We distinguish answers to the survey questions provided by Muslim and Christian participants (i.e. two subcorpora) as a means to examine the discursive choices they make when discussing each other. First, we compare the ICE Nigeria corpus (insofar as the English used by our participants is South West Nigerian English) with our KEO English corpus (e.g. comparison of 'key lemma' lists to explore the 'aboutness' of our corpus and to select a group of words for further study). Second, we compare our two subcorpora (all answers by Muslim and Christian participants). We use a range of corpus outputs for each subcorpus (e.g. word frequencies) to give us an initial insight into the difference and/or lack of difference between the two religious groups under scrutiny, and comment on whether these are meaningful. Third, we delve deeper into the language used by Christians and Muslims to discuss the religious 'other' with a view to gain further insight into what Muslims and Christians' perception of themselves and each other. For example, patterns associated with the words 'Islam' and 'Muslim' on the one hand and 'Christianity' and 'Christian' on the other hand are examined (e.g. using concordance and collocation lists to examine the contexts of these specific words in the two sucbcorpora).

To conclude, our corpus is clearly atypical for it captures data which does not fall neatly under what is usually understood to be a 'corpus' by linguists using a corpus-based paradigm. Our methodological approach therefore raises a range of timely questions and issues for social scientists wishing to use corpus tools in their research. We will therefore also ask what kinds of new lines of enquiry, if any, corpus-assisted discourse analytic methodology can suggest for anthropologists.

Reference

Hunston, S. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

ProtAnt: A freeware tool for automated prototypical text detection

Laurence Anthony
Waseda University
anthony@waseda.jp

Paul Baker
Lancaster University
j.p.baker@lancaster.ac.uk

1 Introduction

Prototypicality can be defined as "having the typical qualities of a particular group or kind of person or thing" (Merriam-Webster 2014). In quantitative and qualitative corpus-based studies, researchers are often interested in identifying prototypical texts so that they can conduct close readings and begin to examine the 'why' behind the numbers revealed through top-down, broad-sweep quantitative analyses. Researchers in other areas may also need to identify prototypical texts in order to, for example, classify texts according to genre, locate typical student essays at a particular level for instructional purposes, flag texts (e.g. extremist writing) for further analysis, or remove outlier texts from a corpus before conducting a quantitative study.

In this paper, we present a novel approach to prototypical text detection that is fast, completely automated, and statistically rigorous. Our method does not require manual assignment of texts to pre-conceived classes as is the case with many natural language processing methods, and it is able to rank texts by their prototypicality in a way that is meaningful and easy to interpret. We have encapsulated our approach in a free software tool, *ProtAnt*, that runs on Windows, Macintosh OS X, and Linux operating systems, and is designed to be easy-to-use and intuitive even for novice users of computers.

2 The *ProtAnt* approach

The starting point for our prototypical text detection approach is to identify key words in a corpus. Key words are 'words' that appear statistically significantly more frequently in the target corpus than in a suitable reference corpus. Depending on the design of the target corpus and choice of reference corpus, these key 'words' may be lexical items, part-of-speech tags, discourse moves, or a multitude of other linguistic features that can be coded or annotated. For this study, we focus on lexical (word) prototypicality. Our *ProtAnt* tool detects these key words using a standard log-likelihood statistically measure of keyness (Dunning 1993), but other measures can be easily incorporated.

The second stage in our approach is to rank the key words so that the most salient key words can be selected for use in further analysis, and the least salient key words removed. Almost all previous corpus-based studies utilizing key words have ranked the words based on the raw 'keyness' value as given by the statistical measure (e.g. log-likelihood). This is equivalent to ranking the words by their p-value. A more informed way to rank key words is by considering the (normalized) size of difference in frequency between the target and reference corpus, i.e., the key word's effect size. There are many ways this can be calculated, including relative frequency (Demarau 1993) or a log of relative frequency (e.g. Hardie 2014). In *ProtAnt*, the user is given a choice of ranking key words by either p-value or effect size measures.

The final stage in the *ProtAnt* approach is to count the number of key words in each corpus file, normalize the counts by the length of the texts, and then rank the corpus texts by the number of key words they contain. Texts containing high numbers of key words are those that contain more words that characterize the corpus as a whole and thus can be considered to be prototypical of the corpus as a whole.

Figure 1 shows a screenshot of the *ProtAnt* tool after completing an analysis of a small corpus of 20 newspaper articles using the BE06 Corpus (Baker 2009) as a reference corpus. In the screenshot, the top right table shows that file 7 is the most prototypical. The middle table shows the key words contained in each file, with file 7 shown to include the words "islam," "blair," "muslim," "brotherhood" and other topic related words. The bottom table shows a complete list of the key words, here created by log-likelihood and ranked by p-values.

3 Validation experiments

Five experiments were conducted to establish the validity of the *ProtAnt* approach to prototypical text identification. The first experiment was designed to see if *ProtAnt* was able to correctly identify prototypical texts in a small corpus of newspaper articles. For this experiment, the corpus was artificially designed to contain 10 texts on the topic of Islam (deemed to be the main theme of the corpus), 5 texts related to the general topic of football (serving as a distractor theme), and 5 texts with no overlapping topics of focus, with the BE06 corpus serving as a reference corpus. A successful *ProtAnt* analysis should be able to rank the 10 texts on Islam higher than the other texts.

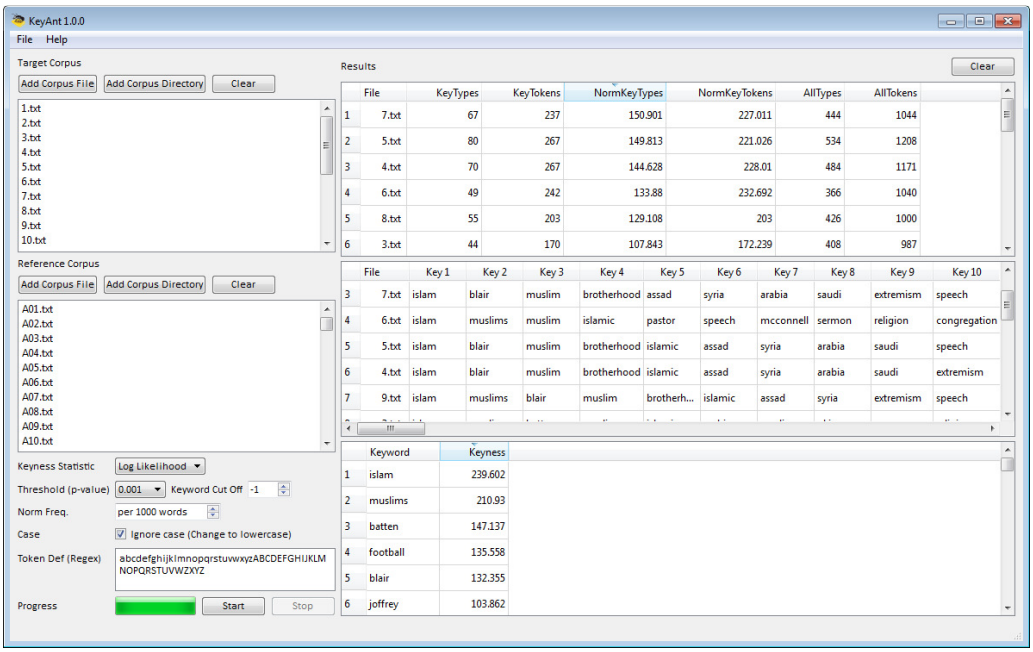


Figure 1: Screenshot of the *ProtAnt* prototypical text detection tool

Table 1 shows the results of the *ProtAnt* analysis for a log-likelihood (LL) threshold value of 0.001 with the texts rank ordered by normalized key type and normalized key token values. Clearly, the *ProtAnt* analysis was able to reliable rank almost all Islam files as the most prototypical of the corpus as a whole, regardless of whether key types or key tokens are used. The rankings were also shown to be stable regardless of the log-likelihood threshold value. Interestingly, one of the Islam texts was unexpectedly ranked lower in the lists. A close reading of this text, however, revealed several unusual features that were not immediately apparent to the investigators; it is a story about a school which told parents that children had to attend a workshop on Islam or be called racist. Thus, this ranking serves as further evidence of the usefulness of the *ProtAnt* tool.

	LL threshold (0.001)	
Rank	Key Types	Key Tokens
1	Islam	Islam
2	Islam	Islam
3	Islam	Islam
4	Islam	Islam
5	Islam	Islam
6	Islam	Islam
7	Islam	Football
8	Islam	Obituary
9	Islam	Islam
10	Football	Islam
11	Obituary	Islam
12	Islam	Football
13	Review	Science
14	Football	Review
15	Science	Islam
16	Tennis	Tennis
17	Football	Football
18	Football	Art
19	Football	Football
20	Art	Football

Table I: *ProtAnt* analysis of newspaper articles

The second experiment was designed to see if *ProtAnt* was able to correctly identify prototypical texts in a small corpus of longer novels. Following a similar design to that used in experiment 1, 10 versions of the novel *Dracula* were compared against five versions of the novel *Frankenstein*, and 5 other randomly selected novels. Again, results revealed that the *ProtAnt* analysis could rank almost

all *Dracula* texts above the other novels in the corpus, with the results remaining stable regardless of key type or key token ordering, or choice of log-likelihood threshold value (results not shown).

Experiments 3 and 4 were designed to see if *ProtAnt* could identify prototypical texts in a larger, traditional corpus. For experiment 3, we performed a *ProtAnt* analysis of texts in the AmE06 Corpus (Potts and Baker 2012) using the BE06 corpus as a reference corpus in order to find prototypical texts that are 'American' in nature. For experiment 4, we performed a *ProtAnt* analysis of texts in the AmE06 Corpus, but this time used the Brown Corpus (Francis & Kucera 1963) as a reference corpus in order to identify prototypical texts expressing the concept of 'the year 2006'. Again, convincing results from the *ProtAnt* analysis were obtained in both experiments, with the highest ranked texts clearly expressing the target themes. For example, in experiment 4, the highest ranked text was a fairly dry government text about tax. It is written with a direct address to the reader and makes frequent use of the second person pronoun key words *you* and *your* (a feature of personalizing language that has become more popular since 1961).

Experiment 5 was designed to see if the *ProtAnt* analysis was able to find outliers in a corpus. For this experiment, we again used AmE06 (with BE06 as the reference), but this time selected all the files from one register and artificially added an additional file randomly selected from a different register. A successful analysis should rank the artificially added file as the lowest in the list. When the experiment was repeated for all registers in AmE06, results showed that the outlier file could be correctly identified as being at the bottom or very close to the bottom of the list (within 2) in 10 out of the 15 cases.

4 Conclusion

In this paper, we have shown that a prototypical text detection approach based on ranking texts according to the number of key words they contain can be successfully applied in a variety of test-case situations. We have also developed a software tool that allows researchers to apply the approach as part of their own analysis through an easy-to-use and intuitive interface. Our software tool, *ProtAnt*, is freely available at the following URL: http://www.laurenceanthony.net/software.html. We hope this tool will introduce traditional qualitative researchers to the advantages of corpus-based approaches, and also remind quantitative corpus-based researchers of the importance of close readings of corpus texts.

References

Baker, P. 2009. “The BE06 Corpus of British English and recent language change”. *International Journal of Corpus Linguistics* 14(3): 312-337.

Damerau, F. J. 1993. “Generating and evaluating domain-oriented multi-word terms from texts”. *Information Processing and Management* 29: 433-447.

Dunning, T. 1993. “Accurate Methods for the Statistics of Surprise and Coincidence”. *Computational Linguistics* 19(1): 61-74.

Francis W. N. and Kucera H. 1964. Brown Corpus. Available online at https://archive.org/details/BrownCorpus

Merriam-Webster. 2014. Available online at http://www.merriam-webster.com/dictionary/prototypical

Potts, A. and Baker. P. 2012. “Does semantic tagging identify cultural change in British and American English?” *International Journal of Corpus Linguistics* 17(3): 295-324.

Tracing verbal *aggression* over time, using the Historical Thesaurus of English

Dawn Archer
University of Central Lancashire
dearcher@uclan.ac.uk

Bethan Malory
University of Central Lancashire
bmccarthy2@uclan.ac.uk

The work reported here seeks to demonstrate that automatic content analysis tools can be used effectively to trace pragmatic phenomena – including *aggression* – over time. In doing so, it builds upon preliminary work conducted by Archer (2014), using Wmatrix (Rayson 2008), in which Archer used six semtags – Q2.2 (speech acts), A5.1+/- ('good/bad' evaluation), A5.2+/- ('true/false' evaluation), E3- ('angry/violent'), S1.2.4+/- ('im/politeness'), and S7.2+/- ('respect/lack of respect') – to examine *aggression* in 200 Old Bailey trial texts covering the decade 1783-93.

Having annotated the aforementioned Old Bailey dataset using Wmatrix, Archer (2014) targeted the utterances captured by the semtags listed above. This afforded her a useful "way in" to (by providing multiple potential indicators of) verbal aggression in the late eighteenth-century English courtroom. Using the 'expand context' facility within Wmatrix, and consulting the original trial transcripts, those incidences identified as verbally aggressive were then re-contextualised – thereby allowing Archer to disregard any that did not point to aggression in the final instance. The success of this approach allowed her to conclude that automatic content analysis tools like USAS can indeed be used to trace pragmatic phenomena (and in historical as well as modern texts).

This approach was not without its teething problems, however. First, apart from those semtags which were used in conjunction with others, as portmanteau tags (e.g. Q2.2 with E3- to capture aggressive speech acts), the approach necessitated the targeting of individual semtags within a given text. The need to perform a time-intensive manual examination of the wider textual context thus made the use of large datasets prohibitive. Furthermore, there was a closely related problem concerning the tagset's basis in *The Longman Lexicon of Contemporary English* (McArthur, 1981), and its consequent inability to take account of diachronic meaning change. This tended to result in the occasional mis-assignment of words which have been subject to significant semantic change over time, including *politely, insult* and *insulted*. In one instance, for example, *politely* was used to describe the deftness with which a thief picked his victim's pocket! The need for manual checks to prevent such mis-assignments from affecting results further necessitated the narrowness of scope to which Archer (2014) was subject.

In the extension to this work, reported here, the authors present their solutions to these problems. These solutions have at their core an innovation which allows historical datasets to be tagged semantically, using themes derived from the Historical Thesaurus of the Oxford English Dictionary (henceforth HTOED). These themes have been identified as part of an AHRC/ESRC funded project entitled "Semantic Annotation and Mark Up for Enhancing Lexical Searches", henceforth SAMUELS[11] (grant reference AH/L010062/1). The SAMUELS project has also enabled researchers from the Universities Glasgow, Lancaster, Huddersfield, Strathclyde and Central Lancashire to work together to develop a semantic annotation tool which, thanks to its advanced disambiguation facility, enables the automatic annotation of words, as well as multi-word units, in historical texts with their precise meanings. This means that pragmatic phenomena such as aggression can be more profitably sought *automatically* following the initial identification of what the authors have termed a 'meaning chain', that is, a series of HTOED-derived 'themes' analogous to *DNA strings*.

This paper reports, first, on the authors' identification of 68 potentially pertinent HTOED 'themes' and, second, on their investigation of the possible permutations of these themes, and the process by which they assessed which themes in which combinations best identified and captured aggression in their four datasets.

The datasets used for this research are drawn from Hansard and from Historic Hansard; and are taken from periods judged to be characterized, in some way, by political/national unrest or disquiet. The datasets represent the periods 1812-14 (i.e., "The War of 1812" between Great Britain and America), 1879-81 (a period of complex wrangling between two English governments and their opposition, led by fierce rivals Disraeli and Gladstone), 1913-19 (the First World War, including its immediate build-up and aftermath), and 1978-9 ("The Winter of Discontent").

References

Archer, D. 2014. "Exploring verbal aggression in English

[11] The SAMUELS project runs from January 2014 to April 2015. For more details, see http://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/samuels/

historical texts using USAS: The possibilities, the problems and potential solutions" In: I. Taavitsainen, A.H. Jucker and J. Tuominen (eds.). *Diachronic Corpus Pragmatics*. Amsterdam: John Benjamins, pp.277-301.

McArthur, T. 1981. *Longman Lexicon of Contemporary English*. London: Longman.

Rayson, P. 2008. "Wmatrix: A Web-based Corpus Processing Environment." Computing Department, Lancaster University. Available online at http://ucrel.lancs.ac.uk/wmatrix/.

Corpus profile of adjectives in Turkish Dictionary (TD): "A" Item Sample

A. Eda Özkan
Mersin University
aedaozkan@mersin.edu.tr

Bülent Özkan
Mersin University
ozkanbulent@mersin.edu.tr

1 Introduction

The researches on the compilation of Turkish Language Vocabulary (TLV) officially started in 1939. These researches are generally based on the collation of previous vocabulary studies. In this sense, the first Turkish Dictionary (TD) was published in 1945. From 1945, there have been 11 editions of Turkish Dictionary (TD) (Türkçe Sözlük, 2010) which contains Turkish Language Vocabulary (TLV). In Turkish Dictionary (TD), containing Turkish Language Vocabulary (TLV), out of 122.423 vocabulary items lexicological defined as utterance, term, idiom, affix and meaning, 12.225 head words are specified as adjective (Türkçe Sözlük, 2010).

TD, containing TLV with its formats in printed, CD and web-enabled, is one of the Turkish Dictionaries which are commonly used and affirmed reliable for today. On the other hand, it is an important point to be indicated that TD is a kind of dictionary which is constituted with lexicological invalid methods. It is known that during the generating process of TD *categorizations, rewriting etc. methods are* used.

Today the principles and methods of corpus linguistics make significant contributions to the studies of lexicology which make researches on the process of forming dictionaries and updating in time through the findings. Today, computational linguistics, also known as *Natural Language Processing* (NLP), by taking the language models called as corpus in parallel to applied linguistics, is commonly used in the studies of lexicology, grammar, dialect, science of translation, historical grammar and linguistic alternation, language teaching and learning, semantics, pragmatics, sociolinguistics, discourse analysis, stylistics and poetics (McEnery et al., 2006: 80-122; Kennedy, 1998: 208-310).

The aforementioned process of forming a dictionary for Turkish, which happens to be corpus based, is a new fact that we have encountered. In this study the contributions of the principles and methods of corpus linguistics to the field in the process of forming TD will be discussed through a sample application.

2 Purpose of the Study

The aim of the study is to present the corpus profile of 812 head words, defined as adjective in A main entry in TD, taken from a corpus of 25 million (+/-) words, formed through the principles and methods of corpus linguistics.

For this object, data set, revealed through a national research project, called "Collocations of Adjectives in Turkey Turkish -A Corpus Based Application-" and supported by TÜBİTAK, will be analyzed.

3 Population – Sample

As a population of this study, *Turkish Corpus - 2* (TC-2), a subject sensitive [art, economy, current news, article, travel, hobby etc.] corpus with 25 million (+/-) words, gathered from various thematic texts which belong to the literary language of Turkish and from the internet environment by using varied software, is used. TC-2, consists of **A-Printed Woks** (%60) between the years of 1923-2008 and **B- Internet Texts** (%40) between the years of 2006-2008.

A- PRINTED WORKS					
	LAYERS	**VARIANCE**		**%**	
1	Novel	96		22,802	
2	Poem	68	7*	16,152	
3	Tale	49	4*	11,638	
4	Essay-Critics	44	3*	10,451	
5	Theatre	35	1*	8,313	
6	Memoir	21		4,988	
7	Research	20		4,750	**%60**
8	Conversation-Interview-Article	18	1*	4,275	
9	Humour	14		3,325	
10	Travel Writing vb.	10	1*	2,375	
11	Letter	4	1*	0,950	
12	Biography	4		0,950	
13	Diary	1		0,237	
14	Various Types	30		7,125	
		403	**18***		
	TOTAL	**421**		**100**	

**Anthological works*

Table 1**.** Content of Corpus

B- INTERNET TEXTS				
	LAYERS	**SUB LAYERS**	**%**	
1	News etc.	*Politics, Economy-Finance, World-Live, Weather Forecast , Sports…*	**%60**	**%40**
2	Life	*Technology, Education, Tabloid Press…*	**%10**	
3	Culture-Art-Health	*Health, Book, Cinema, Theatre…*	**%10**	
4	Essay	*Column…*	**%20**	
			100	
				100

4 Research Questions

According to the corpus queries of 812 lexical items, defined as adjective in "A" item in TD;

- How are the frequency aspects of them?
- How are the use cases of them?
- How are the definition cases of them?
  - The definitions of which adjectives are required to be combined?
  - Which of the adjectives need new definitions?
  - Which adjective's definitions need to be reorganized?

5 Frequency Aspects of Adjectives in TD

The frequenciesy aspects of adjectives (> 100) under "A" in TD are presented in Table 2:[12]

Adjectives	***f***	**Adjectives**	***f***
amansız	*436*	*akli, alışılmış, ani, aylık*	*154*
azgın	*371*	*ayrı*	*153*
ayaklı	*359*	*akıllı, aptal*	*151*
ahşap	*336*	*artçı*	*150*
anlaşılmaz	*307*	*adaletli*	*145*
adlı	*303*	*ağır*	*143*
asırlık	*285*	*ağırbaşlı, analitik*	*140*
aynı	*278*	*alafranga*	*139*
aşırı, ateşli	*272*	*akşamki*	*137*
altmış	*258*	*astronomik*	*134*
apayrı	*244*	*acımasız*	*129*
alaycı	*243*	*altıncı, antidemokratik*	*128*
ailevi	*241*	*anlamsız*	*127*
alt	*239*	*ait, aynalı*	*124*
ayrıcalıklı	*233*	*aklı başında*	*123*
ak	*217*	*aydınlık*	*121*
alaylı	*208*	*acılı*	*118*
altın	*193*	*açık saçık, akademik*	*117*
avantajlı	*185*	*asil*	*116*
agresif	*182*	*abuk sabuk, altın sarısı*	*115*
ahlaklı	*181*	*aptalca*	*113*
arka	*180*	*alakasız, aldatıcı, asabi*	*112*
asılsız	*177*	*akıllıca*	*111*
acıklı, alelade	*175*	*ahenkli, anlamlı*	*110*
arkeolojik, artistik, atıl	*170*	*ağlamaklı, ayrıntılı*	*108*
altı	*163*	*aktif, anlayışlı*	*107*
anlık	*162*	*alımlı*	*104*
acayip, adaletsiz	*160*	*asık*	*103*
Afgan, arsız, asıl, azılı	*157*	*alaturka*	*100*
…	…	…	…

Table 2. The Frequency of Adjectives under "A" Item in TD

[12] The complete list will be presented in main text.

6 The use cases of adjectives in TD

As a result of the corpus queries, the defined adjectives, in TD under "A" item, are presented as the ones in the wild and the ones go out of use in Table 3.

Use	f	%
In use	591	73
Out of use	221	27
Total	**812**	**100**

Table 3. The Use Cases of Adjectives in TD

Table 3 shows that while the usages of 591 adjectives out of the total 812 in TD are seen in the corpus, 221 adjectives are evaluated under the category of "out of use". In this sense, %73 of the analyzed adjectives is formed by the ones in use while %27 of them is defined as out of use.[13]

Among the possible reasons for the case of not being found, although the content of the corpus is a matter of the fact, as in the previous studies on defined lexeme in vocabulary (Özkan, 2010) the reasons such as the lexeme's being old and for this reason their being archaic, belonging to a specific field, being a lexeme from slang or folk speech have impact on the usage of head word lexeme's not being found (Özkan, 2014). The same situation is valid for the lexeme under "A" title in TD.

7 Definition aspects of adjectives in TD

The head word definition aspects of adjectives in TD under "A" item are presented below.

Number of Definition	*f*	Adjective
20	*1*	*ağır*
14	*1*	*açık*
11	*1*	*azgın*
8	*1*	*ayaklı*
5	*4*	*acemi, amatör, anaç, azametli*
4	*7*	*alt, alaylı, aylık, acı, aksak, ak, agresif*
3	*36*	*aynı, ateşli, altın, arsız, alafranga, aydınlık, aktif…*
2	*78*	*alaycı, ailevi, atıl, ani, astronomik, acılı, akademik…*
1	*462*	*amansız, ahşap, anlaşılmaz, adlı, asırlık, altmış, apayrı…*
Total	***591***	**-**

Table 4. The Number of Definitions belong to the Adjectives in TD

On the other hand, as a result of corpus query on the definition aspects of the adjectives under "A" item in TD, it is also possible to analyze these aspects under three titles as: the adjectives, the definitions of which are need to be combined, adjectives which need to have new definitions, and the adjectives, head words definitions of which need to be reorganized.

According to Table 4, among 591 adjectives, one of the adjectives has 20, one of them has 14, one of them has 11, one of them has 8, four of the adjectives have five, seven of them have four, thirty-six of them have three, seventy-eight of them have two and 462 of the adjectives have only one definition.

Adjectives with head word definitions that need to be combined in TD

In TD, the definitions of 33 adjectives under item "A" are need to be combined through the corpus queries as they are not distinctive.[*] These adjectives are listed in Table 5.

Adjectives		
abani	alt	aşırı
afet	amaçlı	aşina
ağırlıklı	anıtsal	aşkın
ahu	anızlı	atıl
ak gözlü	anlayışlı	atılgan
akıl dışı	aptalca	ayaklı
aklı başında	arabalı	aylık
alçak	Arapça	ayrık
aldırmaz	arkasız	ayrıksı
alev kırmızısı	arktik	ayrışık
alkollü	aşınmaz	azgın

Table 5. The Adjectives in TS that Need Definition Combination

Adjectives for which there need to be added new definitions in TD

The results of corpus-based analysis show that 51 adjectives need new definitions. These adjectives are presented in Table 6.

Adjectives		
abdestlik	aksak	arızasız
acemi	alacalı	arkasız
acı	albenili	armut
acımtırak	albenisiz	art
açık	aldatıcı	asık suratlı
adaklı	alt	astarsız
adımlık	altın	aşağılayıcı
agresif	amatör	aşırı
ağdalı	ameliyatlı	atak
ağır	anaç	atılı
ağır yaralı	anadan doğma	avuç
ağırlıklı	angaje	ayaklı
ağızdan dolma	anımsatıcı	aydınlatıcı
ağlı	arabalı	aylık
ağrılı	araçlı	aynı
ağrısız	Arap	ayrıştırıcı
akışkan	arızalı	azgın

Table 6. Adjectives that Need New Definitions TD

[13] The complete list will be presented in main text.

[*] Sample head words will be presented in main text.

Adjectives whose definitions need to be reorganized in TD

Head words definitions 69 adjectives out of 591, found in TD and have usage in TC-2, are required to be reorganized according to definition frequency[*]. These adjectives are shown in Table 7.

abanoz	ak pak	altın	aşağı
acar	akademik	amatör	aşkın
acemi	akıcı	amiyane	atak
acı	akılcı	ampirik	ateşli
acı tatlı	akışkan	anaç	atıl
acılı	aksak	anadan doğma	avuç dolusu
acısız	aksi	ani	ayaklı
açık	aktif	anlamlı	aydın
adsız	alafranga	anonim	aydınlık
ağarık	alaycı	antiseptik	aygın baygın
ağdalı	alaylı	arı	aylı
ağır	albenili	arızalı	aylık
ağırlıklı	alengirli	arızi	aylıklı
ağız dolusu	alevli	aristokrat	aynı
ağrılı	alkollü	arkalı	ayrık
ağrısız	allahsız	arsız	azametli
ahlaksız	alt	astronomik	azgın
ailevi			

Table7. Adjectives, Head Words Definitions of Which are Required to Be Reorganized

8 Conclusion

Every single language goes through changes in terms of vocabulary. Naturally, in every language new words are derived and come into use consistently. In this study lexeme, specified as adjective under "A" item, are investigated through an extensive literary language corpus in terms of frequency, usage, definition frequency and head word definitions.

In the first phase of the study the frequency profiles of the adjectives, which is the research subject, are presented. As its being the primary purpose of lexicography from the beginning to specify the most frequently used lexeme in language and order their definitions, starting from the most frequently used, specification of word class and especially the head words definitions of the ones with type incorporation and the necessity for ordering them on the basis of usage frequency of them (Özkan, 2010), the frequency control is fairly crucial.

As a result of corpus query of adjectives, analyzed in the study, %73 of the adjectives were found to be in use and there was no usage sample for %27 of the adjectives.

During the second phase of the study, the head word definition profiles of the adjectives were analyzed according to many aspects. In this sense, it was observed that a great majority of the adjectives had only one definition.

On the other hand, the study reveals that the definitions of 33 adjectives are need to be combined, 51 adjectives require to be added new definitions and there is a need for the reorganization of the head words definition orders of 69 adjectives (depending on their usage frequency).

The study is significant in terms of presenting the methodological content for the regeneration of a usage-based dictionary of TD in line with the principles of corpus linguistics and lexicology.

Acknowledgement

This study is based upon a National research Project, numbered as TÜBİTAK-SOBAG-109K104 nolu and titled as "Collocations of Adjectives in Turkey Turkish – A Corpus Based Application-". We appreciate the contributions of TÜBİTAK[*].

References

Kennedy, Graeme (1998). *An Introduction to Corpus Linguistics.* New York: Addison Wesley Longman Limited.

McEnery, Tony et al. (2006). *Corpus-Based Language Studies An Advanced Resource Book.* New York: Routledge.

Özkan, B. (2010). "An Investigation on Corpus-Checking of Lexems Defined as "Adverb" in Güncel Türkçe Sözlük." *Turkish Studies International Periodical for the Languages, Literature and History of Turkish or Turkic.* 5/3 Summer 2010: 1764-1782.

Özkan, B. *(2010). Turkish Corpus - 2* (TC-2). Mersin University.

Özkan, B. (2011). TÜBİTAK-SOBAG-109K104 "Collocations of Adjectives in Turkey Turkish - A Corpus Based Application-" Project Report. http://derlem.mersin.edu.tr/ctb/modules.php?module=tanitim

Özkan, B. (2014). "The Corpus-Check of Verbs and the Corpus-Based Dictionary of Verbs in Turkey Turkish Lexicon" *bilig. Journal of Social Sciences of Turkish World.* 69. Spring 2014. 1719-204.

Türkçe Sözlük (2005). Ankara: TDK Yay.

Türkçe Sözlük http://tdk.gov.tr/

[*] The Scientific and Technological Research Council of Turkey.

A corpus-based study of interactional metadiscourse in L1 and L2 academic research articles: writer identity and reader engagement

Juhyun Back
Busan National University of Education

Stance and engagement from the text or the readers are strongly associated with the degree of subjectivity and objectivity in the process of constructing knowledge and developing arguments in academic writing. Of the two main features of metadiscourse, interactional markers play a crucial role in affecting how the writers engage readers or texts and present the writers' authorial identity as scholars in academic community. Although metadiscourse has contributed to the understandings and values of a particular discourse community across different genres in text analysis, cross-sectional variations of interactional metadiscourse by L2 writers in research articles (RA) across different languages and cultures still remain untouched. Few studies have provided answers for how far the L2 writers can understand and use English metadiscourse in the key academic genres which may require particular purposes of the text, the readers, and the social settings. Cross-sectional variations of interactional metadiscourse by L2 writers in research articles (RA) in the field of Applied Linguistics across different languages and cultures still remain untouched. This study analyses interactional metadiscourse used in both NS and NNS research articles (RA) to investigate how far advanced L2 writers can achieve the balance between both objectivity of argumentative writing and professional persona as a scholar within academic community. The investigation comprises several key issues: first, what are the cross-sectional distributions of interactional metadiscourse in both NS and NNS corpus? Second, in what ways may this affect establishing the Korean L2 writer's identity and develop interpersonal relationship within the academic text? Lastly, how can these be explained by their L1 transfer or other developmental factors? To answer the questions, I take a contrastive approach and compare two different sub-corpora from NS (a corpus of 40 research articles written by only English native speakers in the area of Applied Linguistics) and NNS (a corpus of the English research articles of 30 Korean postgraduate students enrolled at doctoral programmes in Korea). The study focuses on the distributions of interactional metadiscourse markers in both NS (181,654 tokens) and NNS (167,905 tokens) corpus, thus identifying the typical features of their linguistic and rhetorical expressions in IMRD structure.

The corpus-based, discourse analytical approach yields several important findings regarding the ways Korean L2 writers at postgraduate level use metadiscourse markers in terms of the establishment of writer identity into the text and of rapport between reader and writer in the genre of research articles. The findings first suggest that NNS showed a higher degree of subjectivity and personality with an overuse of attitude markers, engagement markers, and self-mentions. In the perspectives of stance features, first, Koreans L2 writers may understand the importance of the evaluative function of hedges which express the truth-value of the propositions, as the overall frequency of hedges were in the highest rank in both NNS and NS. However, NNS showed a higher degree of subjectivity and personality with an overuse of interactional metadiscourse markers in several areas: attitude markers, engagement markers, and self-mentions. It is noted that Korean L2 writers tend to present explicit presence of both writer and reader, which may lead to a degree of personality and subjectivity. Such subjectivity may further cause an authoritative voice, with the highest proportion of boosters employed in 'Results' and 'Discussion' section. This failure in keeping objective distance from the text can be partly explained by a socio-pragmatic failure arising from an imperfect understanding of rhetorical traditions across different genres; they may not understand that direct assertiveness is not often welcomed in English academic prose including the genre of research articles. This may not have been transferred from their L1 traditions in that indirectness may be in common in Korean academic discourse.

There were also important cross-sectional differences in self-mentions, attitude markers, and engagement markers. It is notable that the Korean L2 writers at postgraduate level tend to present explicit authorial identities, in particular, conclusions more frequently than in other sections, although they may have been taught to avoid self-mentions, including first-person singular pronouns. This contrasts with the findings from NS corpus in which the higher frequency of self-mentions is shown in 'Introduction' and 'Methods' sections and no serious over-reliance in the frequency of attitude markers occurred in a particular section.

The over-reliance of subjectivity, personality, and engagement features in 'Results' and 'Discussion' section (in particular, in the part of conclusions) may be congruent with the inductive ways of developing organizational patterns in Korean discourse. Despite their lack of linguistic knowledge about the use and functions of metadiscousre markers, this can be partly due to another pragmatic failure transferred

from L1 traditions.

Socio-pragmatic transfer from L1 to L2 may also occur in the Korean L2 writers' preferred linguistic choices in sub-category of each metadiscourse markers; their strong preference for obligation modal verbs among attitude markers might be related to their pragmatic function of hedging in Korean discourse. Also, rhetorical questions among engagement devices, and of modal verb 'would' in hedged expressions, often working as the indirect or politeness discourse strategy in Korean spoken discourse, are more frequently employed by Korean L2 writers. A lack of register awareness might be also problematic. Pedagogical L2 writing resources should be given to teach Korean learners alternative strategies for both genre-specific and culture-specific devices in written academic community.

Sketch Engine for English Language Learning

Vít Baisa
Lexical Computing Ltd., Masaryk Univ
vit.baisa@sketchengine.co.uk

Vít Suchomel
Lexical Computing Ltd, Masaryk Univ
vit.suchemel@sketchengine.co.uk

Adam Kilgarriff
Lexical Computing Ltd.
adam.kilgarriff@sketchengine.co.uk

Miloš Jakubíček
Lexical Computing Ltd. Masaryk Univ
milos.jakubicek@sketchengine.co.uk

There are many websites for language learners: wordreference.com, [14] Using English, [15] and many others. Some of them use corpus tools or corpus data such as Linguee, [16] Wordnik [17] and bab.la. [18] We introduce a novel free web service aimed at teachers and students of English which offers similar functions but is based on a specially prepared corpus suitable for language learners, using fully automated processing, offering the advantages that the corpus is very large – so can offer ample examples for even quite rare words and expressions. We call the service SkELL—Sketch Engine for Language Learning.[19]

SkELL offers three ways for exploring the language:

- Examples: for a given word or phrase up to 40 selected example sentences are shown
- Word sketch, showing typical collocates for the search term
- Similar words, visualized as a word cloud.

Examples (a simplified concordance) is a full-text search tool (see Figure 1).

Word sketches are useful for discovering collocates and for studying the contextual behaviour of words. Collocates of a word are words which occur frequently together with the word—they "co-locate" with the word. For a query, eg *language* (see Figure 2), SkELL will generate several lists containing collocates of the headword mouse. List headers describe what kind of collocates they contain. The collocates are shown in basic word

[14] http://www.wordreference.com
[15] http://usingenglish.com
[16] http://www.linguee.com/
[17] https://www.wordnik.com
[18] http://en.bab.la
[19] While it was tempting to say the 'E' in the acronym should be for 'English', we decided against, as we envisage offering SkELL for other languages (SkELL-it, SKeLL-de, etc).

forms, or 'lemmas'. By clicking on a collocate, a user can see a concordance with highlighted headwords and collocate (using red for the headword and green for the collocate).

The third tool shows words which are similar to a search word, in terms of 'sharing' the same collocates. They may be synonyms, near-synonynms or other related words. For a single word SkELL will return a list of up to forty of the words which are most similar. They are presented as a word cloud (Figure 3).

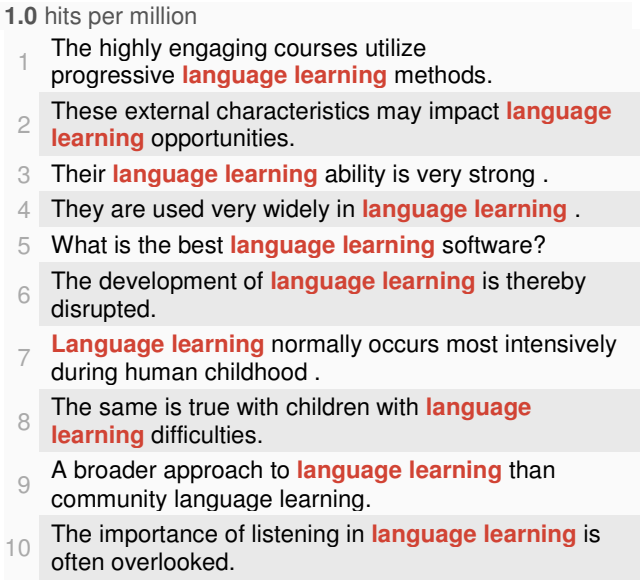


Figure 1: Examples for *language learning*

SkELL uses a large text collection (the 'SkELL corpus') gathered specially for the purpose. We had discovered previously that other collections of over a billion words at our disposal were all from the web, and contained too much spam to use for SKELL: it was critical not to show example sentences to learners if the sentences were not real English at all, but computer-generated junk. The SkELL corpus consists of spam-free texts from news, academic papers, Wikipedia, open-source (non)-fiction books, webpages, discussion forums, blogs etc. There are more than 60 million sentences in the corpus, and one and a half billion words. This volume of data provides a sufficient coverage of everyday, standard, formal and professional English language, even for mid-to-low frequency words and their collocations.

One of the biggest parts of SkELL corpus is English Wikipedia. [20] We included the 130,000 longest articles. Among the longest are articles on *South African labour law, History of Austria, Blockade of Germany:* there are many articles with geographical and historical texts.

Another substantial part consists of books from Project Gutenberg[21]. The largest texts in the PG collection are *The Memoires of Casanova, The Bible (Douay-Rheims version), The King James Bible, Maupassant's Original short stories, Encyclopaedia Britannica.*

We have also prepared two subsets from the enTenTen14, a large general web crawl (Jakubicek et al 2013). The 'White' (bigger) part contains only documents from web domains in www.dmoz.org or in the whitelist of www.urlblacklist.com, as the sites on these lists were known to contain only spam-free material. The 'Superwhite' (smaller) part contained documents from domains listed in the whitelist of www.urlblacklist.com—a subset of White (in case there is still some spam in the larger part taken from www.dmoz.org). The White part contained 1.6 billion tokens.

One part of the SkELL corpus has been built using WebBootCat (Pomikalek et al. 2006, an implementation of BootCaT (Baroni and Bernardini 2004)). This approach uses seed words to prepare queries for commercial search engines.[22] The pages from the search results are downloaded, cleaned and converted to plain text preserving basic structure tags. We assume the search results from the search engine are spam-free, because the search engines take great efforts not to return spam pages to users (wherever there are non-spam pages containing the search terms): BootCaT takes a 'free ride' on the anti-spam work done by the search engines. We have bootcatted approximately 100 million tokens.

We included all of the British National Corpus, as we know it to contain no spam.

The rest of the SkELL corpus consists of free news resources. Table 1 lists the sources used in the SkELL corpus.

Subcorpus	**Tokens** (= words + punctuation) millions	**Tokens used** millions
Wikipedia	1,600	500
Gutenberg	530	200
White	1,600	500
BootCatted	105	all
BNC	112	all
other resources	340	200

Table 1: Sources used for SkELL corpus

As the name says, SkELL builds on the Sketch Engine (Kilgarriff et al 2004), and the corpus was compiled using standard Sketch Engine procedures. We scored all sentences in the corpus using the GDEX tool for finding good dictionary examples (Kilgarriff et al 2008), and re-ordered the whole

[20] https://en.wikipedia.org

[21] https://www.gutenberg.org

[22] We currently use the Bing search engine, http://www.bing.com

corpus so it was sorted according to the score. This was a crucial part of the processing as it speeds up further querying. Instead of sorting good dictionary examples at runtime, all query results for concordance searches are shown in the sorted order without further work needing to be done.

The web interface is available at http://skell.sketchengine.co.uk. There is a version for mobile devices which is optimized for smaller screens and for touch interfaces, available at http://skellm.sketchengine.co.uk.

We have described a new tool which we believe will turn out to be very useful for both teachers and students of English. The processing chain is also ready to be used for other languages. The interface is also directly reusable for other languages, the only prerequisite is the preparation of the specialized corpus. We are gathering feedback from various users and will refine the corpus data and web interface accordingly in the future.

Acknowledgment

This work has been partly supported by the Ministry of Education of the Czech Republic within the LINDAT -Clarin project LM2010013 and by the Czech-Norwegian Research Programme within the HaBiT Project 7F14047.

References

Baroni, M., & Bernardini, S. (2004, May). BootCaT: Bootstrapping Corpora and Terms from the Web. In *LREC*.

Baroni, M., Kilgarriff, A., Pomikálek, J., & Rychlý, P. (2006). WebBootCaT: instant domain-specific corpora to support human translators. In *Proceedings of EAM* (pp. 247-252).

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The TenTen Corpus Family. In *Proc. Int. Conf. on Corpus Linguistics*.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). Itri-04-08 the sketch engine. In *Proceedings of EURALEX* (Vol. 6). Lorient, France. Pp 105-116.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008, July). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of EURALEX* (Vol. 8).

Figure 2: Word sketch for *language*

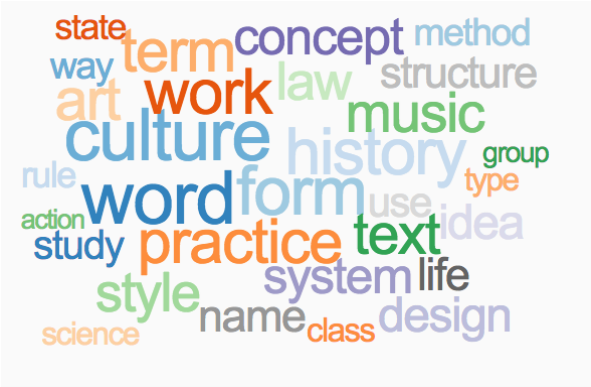


Figure 3: 'Similar words' word clouds for *language* and *lunch*. Size of the word in the word cloud represents similarity to the search word.

Longest-commonest match

Vít Baisa
Lexical Computing Ltd.
Masaryk Univ
vit.baisa@sketchengine.co.uk

Adam Kilgarriff
Lexical Computing Ltd.
adam.kilgarriff@sketchengine.co.uk

Pavel Rychlý
Lexical Computing Ltd.
Masaryk Univ
pavel.rychly@sketchengine.co.uk

Miloš Jakubíček
Lexical Computing Ltd.
Masaryk Univ
Milos.jakubicek@sketchengine.co.uk

1 Introduction

The prospects for automatically identifying two-word multiwords in corpora have been explored in depth, and there are now well-established methods in widespread use. (We use 'multiwords' as a cover-all term to include collocations, colligations, idioms, set phrases etc.) But many multiwords are of more than two words and research into methods for finding items of three and more words has been less successful (as discussed in the penultimate section below).

We present an algorithm for identifying candidate multiwords of more than two words called longest-commonest match.

2 Example

Compare Tables 1 and 2. Both are automatically-generated reports on the collocational behaviour of the English verb *fly*.

Table 2 is an improvement on the input data as shown in Table 1 as it immediately shows:

- two set phrases - as the crow flies, off to a flying start
- sortie occurs as object within the noun phrase operational sorties (a military expression),
- which is generally in the past tense
- flying saucers and insects are salient. The previous level of analysis, in which saucer was
- analysed as object of fly, and insect as subject, left far more work for the analyst to do, including unpacking parsing errors
- sparks go with the base form of the verb

- objects flag and kite, and subjects plane, bird and pilot are regular collocates, occurring in a range of expressions and with a range of forms of the verb.

Gram Relation	Collocate	Freq	Salience
Object			
	saucer	3001	9.92
	kite	376	8.33
	sortie	283	8.17
Subject			
	flag	1176	8.79
	crow	279	8.46
	spark	256	8.02
	aircraft	799	7.84
	plane	527	7.57
	airline	297	7.39
	start	980	7.24
	helicopter	214	7.24
	bird	917	7.08
	insect	245	6.93
	pilot	350	6.68

Table 1: Base collocational data for *fly (v)*

Collocate	Freq	Sal	L-C match	%
saucer	3001	9.92	flying saucers	52.3
flag	1176	8.79	-	
crow	279	8.46	as the crow flies	89.2
kite	376	8.33	-	
sortie	283	8.17	flew operational sorties	47.3
spark	256	8.02	sparks fly	40.6
aircraft	799	7.84	aircraft flying	40.8
plane	527	7.57	-	
airline	297	7.39	airlines fly	30.0
start	980	7.24	off to a flying start	64.8
helicopter	214	7.24	helicopter flying	29.9
bird	917	7.08	-	
insect	245	6.93	flying insects	82.0
pilot	350	6.68	-	

Table 2: As Table 1, but with longest-commonest match. % is the percentage of the hits (column 2) which that the l-c match accounts for.

3 Algorithm

We start from a two-word collocation, as identified using well-established techniques (dependency-parsing, followed by finding high-salience pairs of lexical arguments to a dependency relation.) We then explore whether a high proportion (currently, we use one quarter) of this data is accounted for by a particular string.

The two-word collocations that we start from are triples: <grammatical-relation, lemma1, lemma2>, for example <object, drink_v, tea_n>. The lexical arguments are lemmas, not word forms, and are associated with word class, here represented by underscore and, e. g., *n* for noun, *v* for verb. The corpus instances that will have contributed to giving a high score include "They were drinking tea." and "The tea had been drunk half an hour earlier." The first argument may be to the right of, or to the left of, the second.

If a particular longer string accounts for a high proportion of the data, it becomes a candidate multiword-of-more-than-two-words. We want the string to be common and we want it to be long. Hence the two parts to the algorithm's name. We find the longest-commonest match as follows:

```
Input: two lemmas forming a collocation
candidate, and N hits for the two words

Init: initialize the match as, for each
hit, the string that starts with the
beginning of the first of the two lemmas
and ends with the end of the second.

For each hit, gather the contexts
comprising the match, the preceding three
words (the left context) and the following
three words (the right context)

Count the instances of each match.
Do any of them occur more than N/4 times?
If no, return empty string.
If yes:
       Call this 'l-c match'
       n = Frequency of 'l-c match'
       Look at the first words in its right
and left contexts
       Do any of them occur more than n/4
times?
       If no, return l-c match.
       If yes:
              Take the commonest and add it
       to the l-c match
              Update n to the frequency of
       the new l-c match
              Look at the first words in
       the new right and left contexts
              Do any of them occur more
       than n/4 times?
              If yes, iterate
              If no, return commonest
       extended match
```

We run the algorithm to generate zero, one or more potential extensions of the input to candidate multiwords-with-more-than-one-word.

We currently use a 'one quarter' (n/4) threshold, and a minimum frequency of 5 hits for l-c matches. These were set on the basis of informal reviewing of output. If we can find a more objective way of setting the thresholds, we shall of course do so (and we plan to revise the minimum-frequency threshold so it varies with corpus size).

Note that if there are no common patterns meeting the thresholds for the words between word1 and word2, where they are not adjacent to each other, then there will not be an l-c match.

4 Word forms *vs* lemmas

L-C match also addresses a long-running dispute within corpus linguistics: should collocations be seen as relating to lemmas, or inflected forms? Many prefer lemmas, since it allows more data to be pooled to make generalizations, and if lemmas are not used we are likely to see *invade, invades, invading* and *invaded* in the word sketch for *army*. But others (including many in the 'Birmingham school') object that this moves away from the data and misses critical facts. We are hopeful that the algorithm provides a resolution, presenting constituents of the multiword as lemmas where they occur within the multiword in a range of inflected forms, but as inflected forms, if the multiword generally uses that form.

5 Related work

The paper that opened the field of collocation statistics was Church and Hanks (1989), which introduced (pointwise)[23] Mutual Information as a good statistics for finding two-word collocations in a corpus. Since then work by, inter alia, Evert and Krenn (2001), Wermter and Hahn (2006), Rychlý (2008) has proposed alternative statistics and performed a range of evaluations. Collocation statistics have been integrated into corpus tools and found to be very useful by linguists and lexicographers.

One clear finding is that the route to cleaner collocation lists lies more in the linguistic knowledge applied, in particular, grammar and parsing, then in sophisticated statistics. The way to get the best collocate lists is to apply the best-available grammar and parsing technology for the language: if that is done well, results will tend to be good whatever statistic is used (with sorting according to plain frequency being, for many purposes, as good as or better than any other approach; see also Kilgarriff et al. 2014).

Work that has aimed to extend methods, and statistics, to collocations and other multiwords of more than two words, for example Dias (2003), Daudaravicius and Marcinkeviciene (2004), Petrovic et al. (2010), has had more to say about the statistics than the grammar or parsing. This is unsurprising, since grammar and parsing has not been the research topic of these authors. However it limits the potential that their work has for improving the usefulness of corpus tools, where the hard work lies in the language-specific POS-tagging, lemmatization, grammar and parsing.

One problem for finding general, language-independent, corpus-independent solutions is that languages (and the technologies available for them) vary across languages: another is that corpora vary in size by many orders of magnitude, and, within a corpus, word- and multiword-frequencies also vary by orders of magnitude. Statistics that work well for a 100,000-word corpus may or may not make sense for a 10-billion-word corpus. Statistics that work well to find idioms with corpus frequencies of ten or less may not work well at identifying colligational patterns with frequencies in the millions.

While evaluation exercises are of value, they are usually based on a single corpus, language and target multiword-type. We should not over-generalize.

In our approach we aim to make the maximum benefit of dependency parsing, specially for the two-word case where we know it works well, and to use simple methods, which are, we hope, fairly scale- and language-independent, to build on what that gives us. We make substantial efforts to make high-quality dependency parsing (as well as lemmatization and POS-tagging) available in our system, the Sketch Engine (Kilgarriff et al 2004), for a large number of languages.

6 Current status

An earlier version of the longest-commonest algorithm was already presented in Kilgarriff et al (2012). We (re-)present the work because it was only covered very briefly in the earlier presentation, and in the meantime we have developed a version of the algorithm that works very fast even for multi-billion word corpora, and is fully integrated into our corpus query system, the Sketch Engine.

Acknowledgement

This work has been partly supported by the Ministry of Education of the Czech Republic within the LINDAT-Clarin project LM2010013 and by the Czech-Norwegian Research Programme within the HaBiT Project 7F14047.

References

Church, K. W., Hanks, P. 1989. Word association norms, mutual information, and lexicography. Proc 27th ACL, Vancouver, Canada. Pp. 76–83.

Daudaravičius, V., Marcinkevičienė, R. 2004. Gravity counts for the boundaries of collocations. Int Jnl of Corpus Linguistics 9(2) pp. 321–348.

[23] While Church and Hanks call their statistic 'mutual information', it has been pointed out since that this is not the standard usage in the information-theory literature, and their statistic is usually called 'pointwise mutual information'.

Dias, G. 2003. Multiword unit hybrid extraction. Proc. ACL workshop on Multiword expressions: analysis, acquisition and treatment. Pp 41–48.

Evert, S., Krenn, B. 2001. Methods for the qualitative evaluation of lexical association measures. Proc 39th ACL, Toulouse, France. Pp. 188–195.

Kilgarriff, A., Rychlý, P., Smrz, P., Tugwell, D. 2004. The Sketch Engine. Proc. EURALEX. pp. 105–116.

Kilgarriff, A., Rychlý, P., Kovář, V., Baisa, V. 2012. Finding multiword of more than two words. Proc. EURALEX. Oslo, Norway.

Kilgarriff, A., Rychlý, P., Jakubicek. M., Kovář, V., Baisa, V. and Kocincová, L. 2014. Extrinsic Corpus Evaluation with a Collocation Dictionary Task. Proc LREC, Rejkyavik, Iceland.

Petrovic, S., Snajder, J., Basic, B.D. 2010. Extending lexical association measures for collocation extraction. Computer Speech & Language 24(2) pp. 383–394.

Rychlý, P. 2008. A Lexicographer-Friendly Association Score. Proc. RASLAN workshop, Brno, Czech Republic.

Wermter, J., Hahn, U. 2006. You can't beat frequency (unless you use linguistic knowledge) – a qualitative evaluation of association measures for collocation and term extraction. Proc. 44th ACL, Sydney, Australia. Pp. 785–792.

Triangulating methodological approaches (*panel*)

Paul Baker
Lancaster University
p.baker@lancaster.ac.uk

Jesse Egbert
Brigham Young University
jesse_egbert@byu.edu

Tony McEnery
Lancaster University
a.mcenery@lancaster.ac.uk

Amanda Potts
Lancaster University
a.potts@lancaster.ac.uk

Bethany Gray
Iowa State University
begray@iastate.edu

1 Introduction

This panel is based on a forthcoming edited collection which aims to 1) Showcase a wide variety of corpus linguistic methods through a series of parallel empirical studies using a single corpus dataset; 2) Investigate the extent to which these different methods can complement one another by examining similarities and differences among the findings from each approach; and 3) Explore the potential for future triangulation of corpus methodologies in order to enhance our understanding of corpus data.

More specifically, we have given the same corpus to ten analysts and asked them to analyse it independently of one another. The panel will first discuss the rationale for the project and describe the corpus. Following this, individual panelists will present some of their findings using the various methods adopted. The final part of the panel will involve a comparison of the methods and a discussion of the extent to which (different forms of) triangulation are likely to result in favourable research outcomes. We aim to assess the different methods in relation to each other to determine the extent to which they have "complementary strengths and nonoverlapping weaknesses" (Johnson et al. 2007).

2 Triangulation

Methodological triangulation has been used for decades by social scientists as a means of explaining behavior by studying it from two or more perspectives (Cohen & Manion 2000: 254). Contemporary corpus linguists use a wide variety of methods and tools to study linguistic and discursive

patterns. However, only a small amount of research has been carried out on triangulation in corpus linguistics and has tended to use small numbers of analysts. For example, Baker (2015) involved 5 analysts carrying out a critical discourse analysis of a corpus of newspaper articles while Baker (2014) involved the author comparing three methods. Marchi and Taylor (2012) have also carried out triangulation experiments which involved the two analysts separately conducting analyses of the same corpus. Our project seeks to carry out a fuller investigation of a wider range of techniques and analyst perspectives on a single corpus. We are particularly interested in the extent to which the findings from the different investigations are similar, complementary (e.g. different but not contradictory) or divergent.

3 The Q+A Corpus

We have compiled a 400,000 word corpus consisting of web pages from online question and answer (Q+A) forums. Based on a random sample of 1,000 internet pages we estimate that such pages make up about 11% of web pages. Computer mediated communication is an ideal register to focus our project around as it (a) is largely unexplored by corpus linguists, (b) enables examination of linguistic innovation, and (c) lends itself to comparisons with spoken and written registers. Also, the Q+A forums often involve discussion of topics such as current events, relationships, religion, language, family, society and culture which are pertinent to discourse analytical approaches.

Data was collected from equivalent sites for the UK, the US, India and the Philippines, allowing the potential for comparison between different Englishes. The texts in the corpus were balanced across 3 general topic areas (society & culture, family & relationships, and politics & government). Each text was annotated with tags marking the start and end of (a) the question, (b) each answer, and (c) the best answer. Additionally, a version of the corpus was grammatically annotated using the CLAWS C7 annotation system.

4 Research question

Each analyst was asked to use a particular method of their choice in order to answer the following, purposefully broad research question: 'In what ways does language use in online Q+A forum responses differ across four world English varieties (India, Philippines, United Kingdom, and United States) and/or among topics (society & culture, family & relationships, and politics & government).

5 Analytical methods

The following ten analytical methods have been undertaken by different authors:

Keyword analysis – a corpus driven approach which compares each of the four different language varieties against the others in order to identify words which are statistically more frequent in one part of the corpus when compared against the remainder. Keywords ought to identify words are phrases that may be specific to individual topics or questions asked, but they may also be revealing of authorial style, and if used by multiple authors in the corpus, could identify language features associated with a particular regional register.

Semantic field analysis – to describe and compare the various subcorpora of online question and answer forum texts. This is achieved through automated semantic tagging and calculation of statistical significance using USAS and Wmatrix. The use of semantic categories is particularly helpful when comparing subcorpora of relatively small sizes, as with this dataset. Rather than restricting one's view to the word level (where infrequent words will be disadvantaged), it is possible to group many types with similar meanings together, allowing for analysis of a greater variety of features with significant frequency.

Lexical bundles in the Q+A corpus will be analyzed along three major parameters: (a) linguistic form, including their structural composition and degree of fixedness; (b) discourse function, aligning with the established stance-conveying, discourse-organizing, and referential functions of lexical bundles as well as allowing for the identification of discourse functions specific to Q+A forums; and (c) the distribution of the bundles across the corpus.

Multifactorial approaches to variation in lexis/syntax. Using as an example the variation in choices of future marking (*will* vs. *going to* vs. *shall*) this approach examines how different linguistic and contextual factors affect the choice of future marking in an internet-based corpus covering different varieties of English and how regression methods can help shed light on potentially complex interactions between multiple predictors of future choice.

Multi-Dimensional Analysis is based on the theoretical assumption that functional dimensions of texts involve underlying patterns of linguistic co-occurrence. This approach to linguistic variation suggests that systematic differences may occur in a corpus of Q+A forums as writers make lexical and grammatical choices appropriate to this register. MD data are obtained from factor analysis which considers the sequential, partial, and observed correlations of a wide-range of variables producing

groups of statistically co-occurring features.

Stylistic Perception Analysis is a new method of investigating linguistic variation from the perspective of audience perceptions. In SP analysis multiple participants are asked to read each text sample in a corpus and respond to a series of semantic differential items designed to measure their perceptions of the style of the text (e.g., quality, readability, relevance). Correlations can then be explored between linguistic variation and reader perceptions of writing style.

Pragmatics deals with implied and inferred meanings, with intentional and unintentional meanings, with the dynamic and the emergent – in short, with phenomena that leave little overt trace in the text, or are even entirely "invisible". Concepts such as annotation, collocation, and lexical bundles can help us study (1) speech acts, speech act sequences and speech act frames, (2) forms conventionally enriched with pragmatic meanings, and (3) meta-pragmatic comments and labels, revealing, for example, psychological states and attitudes.

Corpus-assisted Discourse Analysis of gendered discourses. Frequency and concordance analyses of a set of words relating to gender (e.g. *man*, *woman*, *male*, *herself* etc.) are carried out in order to identify variation in gendered discourses. Two forms of comparison are made – a) a sex-based comparison which focusses on identifying gendered discourses which position men and women differently or similarly and b) a cultural-based comparison which compares gendered discourses across UK, US, India and the Philippines.

Collocational network analysis. Using the tool GraphColl, which plots visual representations of collocational relationships, a study of a small number of highly frequent words across all four registers will be carried out, in order to identify how such words may occur in different contexts, depending on the register they occur in.

Qualitative discourse analysis. Using a tool called ProtAnt, which uses the keywords technique in order to identify files in a corpus which are most typical or central in terms of key lexis, the corpus data set is narrowed down to just three files from each of the four language varieties. These files are given to an analyst who then conducts a qualitative 'close reading' of them, without using any corpus tools or techniques.

References

Baker, P. (2014). *Using Corpora to Analyse Gender*. London: Bloomsbury.

Baker, P. (2015). Does Britain need any more foreign doctors? Inter-analyst consistency and corpus-assisted (critical) discourse analysis. In M. Charles, N. Groom and S. John (eds) *Grammar, Text and Discourse.* Amsterdam/Philadelphia: John Benjamins.

Cohen, L., & Manion, L. (2000). *Research methods in education.* London: Routledge.

Johnson, R.B., Onwuegbuzie, A.J., & Turner, L.A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research,* 1(2): 112 – 133.

Marchi, A. & Taylor, C. 2009. If on a Winter's Night Two Researchers... A challenge to assumptions of soundness of interpretation. *Critical Approaches to Discourse Analysis across Disciplines* 3(1): 1–20.

Gender distinctions in the units of spoken discourse

Michael Barlow
University of Auckland
mi.barlow@auckland.ac.nz

Vaclav Brezina
Lancaster University
v.brezina@lancaster.ac.uk

1 Introduction

Since there are various commonly-held assumptions about differences between men and women, the use of empirical data is especially important in determining the nature and extent of differences in spoken usage between the gender groups. So far, corpus-based sociolinguistic studies have typically offered general comparison of frequencies of a target linguistic variable in socially defined sub-corpora (e.g. speech of all men vs. speech of all women in the corpus). This procedure, however, emphasises inter-group differences and ignores within-group variation because most of these studies do not use any measure of dispersion (cf. Brezina & Meyerhoff 2014; Gries, 2006). The research reported here differs from much of the previous corpus-based gender research by examining the frequency of linguistic elements in particular positions in the utterance: initial, final, medial etc. The selection of an appropriate corpus for research on gender requires some thought because of the problem of confounding variables. The approach taken in this investigation is explained in the following section.

2 Corpus data

The study is based on *BNC 64*, a 1.5-million-word corpus of casual speech extracted from the *BNC – demographic*. *BNC 64* is a corpus which represents the speech of 64 selected speakers (32 men and 32 women) who provide between 6.4 and 64 thousand tokens each. In addition to gender, the corpus is also balanced for age, socio-economic status and region (see Table 1). In *BNC 64,* the transcribed speech from each individual speaker is stored in a separate file which enables us to easily explore both individual and social variation in the corpus.

Gender	Age	Socio-econ. status		Region
32 M 32 F	A (14-34): 24 B (35-54): 27 C (55+): 13	AB: 14 C1: 16	MC: 30	different regions across the UK
		C2: 17 DE: 13	WC: 30	
		UU: 4		

Table 1: BNC 64 - Basic characteristics

3 Method

The aim of the study is to examine the frequency of occurrence of words and phrases in different positions in the utterance. This is accomplished using a software program, *WordSkew* (Barlow, 2014).

WordSkew allows the user to determine the frequency of words, phrases or POS tags across portions of different textual units: sentences, paragraphs, or the complete text. The "portions" can either be calculated in relation to equal division of the unit --- first 10%, second 10%, etc --- or as absolute positions such as first word in the sentence, first sentence in the paragraph etc. Thus it is possible to search for positions "1 2 Mid -1 #" where # stands for last position and -1 is the penultimate position. The software gives the results as histograms (and tables). Clicking on a particular bar of the histogram reveals the concordance lines for that position. For the current investigation, the textual unit is the utterance because we are dealing with spoken data.

The research reported here is exploratory and is not testing any particular theoretical stance. Some general search probes are used to investigate potential differences between men and women, while also keeping track of individual differences.

4 Results

The following graph (Figure 1) shows the use of common sentence-initial bigrams by male and female speakers. Because *BNC64* allows us to trace the use of linguistic variables by individual speakers, we present the means for gender groups with error bars (showing 95% confidence intervals), which reflect internal variation within each gender group.

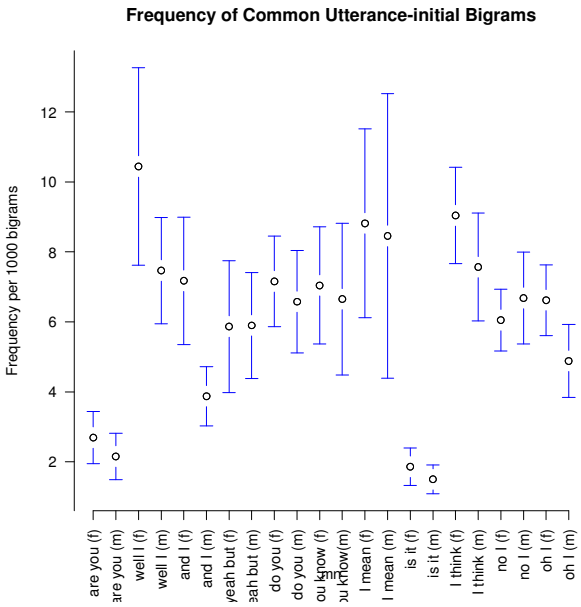


Figure 1: F/M bigrams in utterance-initial position

The details of the results may be difficult to discern in this graph. Some overall patterns of similarity and difference can, however, be observed: The left two data points (with confidence intervals) show the low frequency of use of *are you* in initial position, with women on average using it slightly more than men. The next pair of data points reveal a more marked preference by women for utterance-initial *well I*. The next bigram data is *and I* which is also strongly preferred by women. One bigram preferred by male speakers is *no I*.

When we look at the utterance-second position, we can see a preference for a different set of bigrams as well as a considerable amount of individual variation (indicated by long error bars) through which patterns of gender variation emerge (Figure 2).

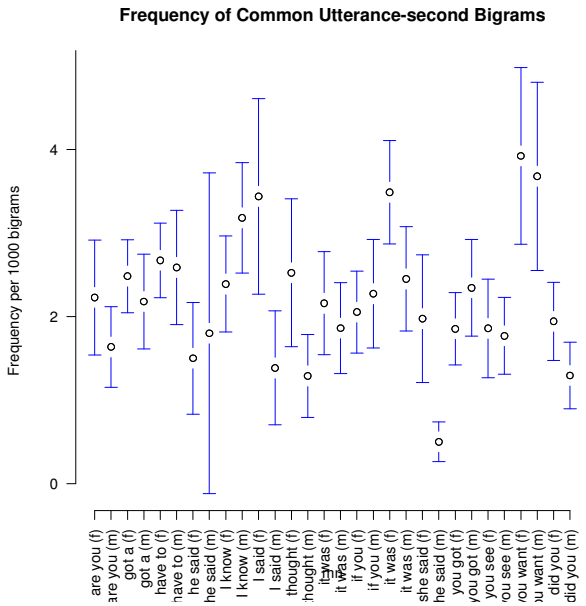


Figure 2: F/M bigrams in utterance-second position

Again we can pick out some of the results from the data in Figure 2. Male speakers have a preference for *I know* in this position compared with female speakers. In contrast, *I said* was more frequently used by women than men.

With *WordSkew* we can also trace a single linguistic feature across utterance positions as is demonstrated in Figure 3 which shows the use of *I know* and *you know* in the first, second and utterance-last positions.

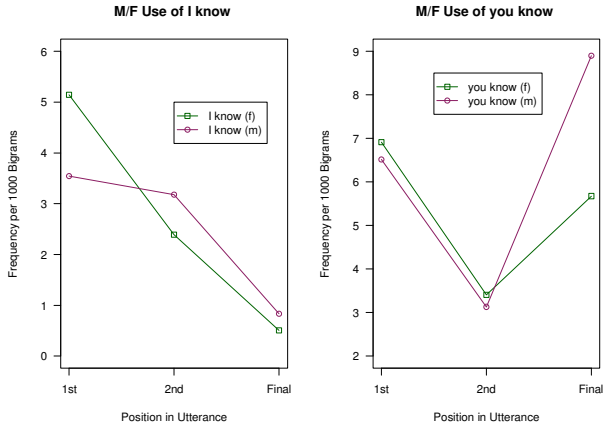


Figure 3: M/F differences in frequency of use of *I know* and *you know* by position

Although the frequencies are rather low, we see some suggestive patterns in the data, pointing to interesting differences in the speech of men and women. Overall, we see very similar usage. However, there is an indication of a preference for use of *I know* by women in initial position and a preference for *you know* by men in final position.

5 Discussion

The results are necessarily preliminary since they are based on one fairly small corpus. However, there are two notable aspects to the current research. One is based on the results related to the position of linguistic features in an utterance. The differences in usage by men and women are tied to particular positions of the linguistic features in the utterance. Figures 1 and 2 can be analysed to find data of interest for further investigation. The next step is to extract specific patterns as illustrated in Figure 3. Thus we propose a more refined methodology than one based on comparative frequency without reference to position.

A second aspect of the research relates to individual variation in the data. Although many investigations compare overall frequency of use by speaker groups, we propose a more refined methodology which includes building confidence intervals around means so that the extent of individual variation can be assessed.

References

Barlow, M. (2014) *WordSkew,* Athelstan: Houston.

Brezina, V., & Meyerhoff, M. (2014). "Significant or random?: a critical review of sociolinguistic generalisations based on large corpora". *International Journal of Corpus Linguistics*, *19*(1), 1-28.

Gries, S. T. (2006). Exploring variability within and between corpora: some methodological considerations. *Corpora, 1*(2), 109-151.

"All the news that's fit to share": Investigating the language of "most shared" news stories

Monika Bednarek
University of Sydney
Monika.Bednarek@sydney.edu.au

James Curran
University of Sydney
james.r.curran@sydney.edu.au

Tim Dwyer
University of Sydney
timothy.dwyer@sydney.edu.au

Fiona Martin
University of Sydney
fiona.martin@sydney.edu.au

Joel Nothman
University of Sydney
joel.nothman@gmail.com

1 Introduction

The sharing of news via social media services is now a significant part of mainstream online media use and is an increasingly important consideration in journalism practice and production. This project analyses the linguistic characteristics of online news-sharing on Facebook. It is part of a larger, multidisciplinary project funded by the Australian Research Council (ARC LP140100148) which brings together methods from computing science, linguistics and audience research with the aim of developing an analytical framework for monitoring, classifying and interpreting news-sharing practices that can inform media industry development, journalism education and digital media policy. The project team includes researchers in Journalism Studies, Information Technologies and Linguistics, working in collaboration with Australian media industry partners Mi9 and Share Wars.

2 Corpus

As a first case study, we compiled a corpus of the top 100 'most shared' news stories. Our aim was to establish a base-line by examining first those most-shared news stories that originate with print and broadcast English-language 'heritage' news media organisations (such as *New York Times, Guardian, CNN*) rather than 'digital natives' (new media organisations such as *Buzzfeed, Upworthy, Huffington Post*). The business model of these publishers focuses on promoting news sharing and they employ a greater array of techniques to encourage this behaviour. Therefore a baseline study will help to identify any differences between old and new media. We also excluded magazines (such as the *Atlantic*). To compile the corpus, we used ShareWar's Likeable Engine[24] to extract the top 200 items by total Facebook share count as at early September 2014. We then manually excluded any items that were not news stories, for example quizzes, advice, online picture galleries, videos, or opinion. The final corpus contains the 100 news items from English-language news media organisations that were the most successful in terms of their Facebook share count. The decision to start with a small corpus of 100 stories was deliberate, as it allows us to combine quantitative and qualitative corpus and discourse analytical techniques, which will inform later analyses of larger corpora – including a comparison with news stories distributed by new media organisations such as *Huffington Post* and *Buzzfeed*.

3 Analyses

Analyses combine the application of classic corpus linguistic tools (such as frequency and keyness analysis as well as concordancing) with manual, computer-aided annotation. The main focus of the analyses is on discursive news values analysis (DNVA), as developed by Bednarek & Caple (2012a, b). This type of analysis focuses on newsworthiness, i.e. the worth of a happening or issue to be reported as news, as established via a set of news values (such as Negativity, Proximity, Eliteness, Unexpectedness, etc). Discursive news values analysis examines how this 'worth' – and these news values – are established through semiotic resources and practices. This project focuses on linguistic rather than other semiotic resources. DNVA can proceed via 'manual' close-reading discourse analysis and/or via the use of automatic corpus techniques. A corpus linguistic approach has only been employed in three previous DNVA studies:

Bednarek and Caple (2012b) use frequency lists and concordancing for analysis of news values in *one* environmental news story, complementing this with manual multimodal discourse analysis. Bednarek and Caple (2014) suggest that various corpus linguistic techniques can be used to study newsworthiness. However, for reasons of scope, they focus only on word/bigram frequency and keywords, applying two different methods to a small corpus (approximately 70,000 words): The first method is to manually identify, from frequency/keywords lists, those forms that seem to have the potential to construct news values. These are called "pointers" (Bednarek and Caple, 2014: 145) to newsworthiness. The second method is to investigate topic-associated words using

[24] http://likeable.share-wars.com/

concordancing to gain insights into which news values are associated with particular concepts or entities.

Potts, Bednarek and Caple (in press) use a 36-million word corpus of news reporting on Hurricane Katrina in the US to explore how computer-based methods can help researchers to investigate the construction of newsworthiness. They test and evaluate the integration of corpus techniques in applying discursive news values analysis (DNVA). These techniques include tagged lemma frequencies, collocation, key POS tags and key semantic tags.

This case study builds on these studies, but has a more focussed research question: what kinds of news values are emphasized in 'most shared' news stories of legacy media and how are these values constructed linguistically? This will provide a baseline for understanding the construction of those values in a more diverse corpus which includes stories from digital natives. Results contribute to urgently needed knowledge about the meaning and consequences of changing modes of news dissemination, addressing a key concern for industry development, journalism practice and media studies as online media markets expand: what factors shape news-sharing on social media?

Acknowledgments

This paper is an output of the Australian Research Council Linkage Project grant *Sharing News Online: Analysing the Significance of a Social Media Phenomenon* [LP 140100148].

References

Bednarek, M. and Caple, H. 2012a. *News discourse*. London/New York: Continuum.

Bednarek, M. and Caple, H. 2012b. '''Value Added': Language, image and news value". *Discourse, Context & Media* 1: 103-113.

Bednarek, M. and Caple, H. 2014. "Why do news values matter? Towards a new methodological framework for analyzing news discourse in Critical Discourse Analysis and beyond". *Discourse & Society* 25 (2): 135-158.

Potts, A., Bednarek, M. and Caple, H. in press. "How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina". *Discourse & Communication.*

Identifying linguistic epicentres empirically: the case of South Asian Englishes

Tobias Bernaisch
Justus Liebig University Giessen
Tobias.J.Bernaisch@anglistik.uni-giessen.de

Stefan Th. Gries
University of California, Santa Barbara
stgries@linguistics.ucsb.edu

A linguistic epicentre can generally be identified on the basis of two criteria, namely "it shows endonormative stabilization (i.e. widespread use, general acceptance and codification of the local norms of English) [...] on the one hand, and the potential to serve as a model of English for (neighbouring?) countries on the other hand" (Hundt 2013: 185). Along these lines, "(pre-) epicentric influence" (Peters 2009: 122) has been traced from Australian on New Zealand English and Leitner (cf. 1992: 225) posits that Indian English is a linguistic epicentre for South Asia. Studies on epicentral variety constellations, however, have so far mainly explored "degrees of similarity between a specific dominant variety on the one hand (i.e. BrE [= British English] or Indian English) and peripheral varieties on the other (e.g. Sri Lankan English and Pakistani English)" (Hoffmann et al. 2011: 261) based on whether statistically significant differences in surface structure frequencies of a given phenomenon exist or not. Given that these studies did not investigate the underlying variety-specific norm-related models triggering these surface structure choices, it seems that analyses of potential epicentral configurations "still lack the empirical evidence that would allow us to make more than educated guesses" (Hundt 2013: 186). This study suggests an empirical, corpus-based and model-oriented method of epicentre identification and applies it to South Asian Englishes.

With a focus on the dative alternation, i.e. the alternation between the double-object construction (e.g. *John gave Mary a book.*) and the prepositional dative (e.g. *John gave a book to Mary.*), the norms underlying this constructional choice are studied in six South Asian Englishes and British English. The corpus data for South Asia stem from the 18m-word-large *South Asian Varieties of English* (SAVE) *Corpus* (cf. Bernaisch 2011) sampling variety-specific acrolectal newspapers texts of Bangladeshi, Indian, Maldivian, Nepali, Pakistani and Sri Lankan English and the newspaper texts in the *British National Corpus* are used as British English equivalents. Via *Multifactorial Prediction and*

Deviation Analysis with Regression (MuPDAR; cf. e.g. Gries and Adelman 2014) under consideration of nested random effects, we a) identify the factors that cause South Asian speakers of English to make constructional choices different from British English speakers when other influences on the constructional choice are controlled for and b) the South Asian linguistic epicentre in a completely bottom-up fashion by empirically validating the linguistic model which best represents the norms underlying the dative alternation in South Asian Englishes.

1381 examples were – under consideration of earlier findings on the dative alternation (cf. e.g. Gries 2003, Bresnan and Hay 2008, Schilk et al. 2013, Bernaisch et al. 2014) – annotated for syntactic (the verb-complementational pattern used, length and pronominality of patient and recipient), semantic (animacy of patient and recipient, the semantic class of the ditransitive verb), pragmatic (the discourse accessibility of patient and recipient) and data-structure-related variables (the newspaper from which a given example from a particular variety was taken). In terms of differences between British English and South Asian speakers of English, the results *inter alia* show that speakers of South Asian Englishes choose more prepositional datives than British English speakers when the patient or the recipient are not introduced in the preceding discourse and when there is no marked difference in the lengths of patient and recipient. Example (1) taken from the Bangladeshi SAVE component illustrates this.

(1) Of course, I am not proposing that we should give the valuable space of Daily Star to fascists like Falwell. <SAVE-BAN-DS_2003-06_pt29>

Triggered by the newly introduced recipient "fascists like Falwell", the Bangladeshi speaker here opts for a prepositional dative. In British English, new recipients also prefer prepositional datives, but due to other characteristics of the verb phrase, British English speakers are predicted to choose a double-object construction in this example. In the light of this, it seems to be the case that the cue 'new recipient' for prepositional datives is notably stronger in South Asian Englishes than in British English. Given that similar observations can be made for new patients, discourse accessibility of patient and recipient seems to be an actuator of structural nativisation (cf. Schneider 2003, 2007) in South Asian Englishes. In a second step, the linguistic epicentre of South Asia is identified by analysing the norm-related models guiding the constructional choices of the dative alternation in the varieties under scrutiny. Based on iteratively comparing how well a variety-specific model derived via MuPDAR can predict constructional choices in the remaining varieties, we are able to show that it is valid to assume that Indian English functions as a linguistic epicentre for South Asia – at least in relation to the dative alternation. Given that Indian English can be regarded as an endonormatively stabilised variety (cf. Mukherjee 2007: 163), this finding is certainly in accordance with the advanced evolutionary status linguistic epicentres should theoretically display (cf. Hundt 2013: 185) and provides strictly empirical evidence for earlier, partly introspective perspectives on epicentral configurations in South Asia (cf. e.g. Leitner 1992).

References

Bernaisch, T., Gries, S.T. and Mukherjee, J. 2014. "The dative alternation in South Asian English(es): modelling predictors and predicting prototypes". *English World-Wide* 35(1): 7–31.

Bernaisch, T., Koch, C., Mukherjee, J. and Schilk, M. 2011. *Manual for the* South Asian Varieties of English *(SAVE)* Corpus*: compilation, cleanup process, and details on the individual components*. Giessen: Justus Liebig University.

Bresnan, J. and Hay, J. 2008. "Gradient grammar: an effect of animacy on the syntax of *give* in New Zealand and American English". *Lingua* 118: 245–259.

Gries, S.T. 2003. "Towards a corpus-based identification of prototypical instances of constructions". *Annual Review of Cognitive Linguistics* 1: 1–27.

Gries, S.T. and Adelman, A.S. 2014. "Subject realization in Japanese conversation by native and non-native speakers: exemplifying a new paradigm for learner corpus research". In J. Romero-Trillo (ed.) *Yearbook of Corpus Linguistics and Pragmatics 2014: New Empirical and Theoretical Paradigms*. Cham: Springer.

Hoffmann, S., Hundt, M. and Mukherjee, J. 2011. "Indian English – an emerging epicentre? A pilot study on light verbs in web-derived corpora of South Asian Englishes". *Anglia* 129(3–4): 258–280.

Hundt, M. 2013. "The diversification of English: old, new and emerging epicentres". In D. Schreier and M. Hundt (eds.) *English as a Contact Language*. Cambridge: Cambridge University Press.

Leitner, G. 1992. "English as a pluricentric language". In M. Clyne (ed.) *Pluricentric Languages: Differing Norms in Different Nations*. Berlin: Mouton de Gruyter.

Mukherjee, J. 2007. "Steady states in the evolution of New Englishes: present-day Indian English as an equilibrium". *Journal of English Linguistics* 35(2): 157–187.

Peters, P. 2009. "Australian English as a regional

epicentre". In T. Hoffmann and L. Siebers (eds.) *World Englishes – Problems, Properties and Prospects*. Amsterdam/Philadelphia: John Benjamins.

Schilk, M., Mukherjee, J., Nam, C.F.H. and Mukherjee, S. 2013. "Complementation of ditransitive verbs in South Asian Englishes: a multifactorial analysis". *Corpus Linguistics and Linguistic Theory* 9(2): 187–225.

Schneider, E.W. 2003. "The dynamics of New Englishes: from identity construction to dialect birth". *Language* 79(2): 233–281.

Schneider, E.W. 2007. *Postcolonial English: varieties around the world*. Cambridge: Cambridge University Press.

'May God bless America': Patterns of in/stability in presidential discourse

Cinzia Bevitori
University of Bologna

cinzia.bevitori@unibo.it

1 Introduction: Aims and purpose

This paper builds on an ongoing research project aiming to explore diachronic language variation in specialised corpora of U.S. Presidential speeches (see Bayley and Bevitori 2011, 2014, 2015; Bevitori 2014; 2015). In particular, the analysis will focus on a corpus consisting of a complete set of transcripts of State of the Union addresses delivered by U.S. Presidents and covering a span of time of more than two hundred years, from 1790 to 2014. As a crucial presidential moment, in fact, the State of the Union address stands out as a symbol and instrument of 'national unity'. The diachronic specialized corpus has thus been build in order to be representative of this register of discourse and, as such, it may provide a valuable resource for investigating language change (and continuity) and its interplay with politics in one of the most powerful settings of institutional discourse: the American Presidency.
The aim of this study is threefold. First, it aims at investigating how and to what extent religion intersects with politics within this highly specialized political domain. In fact, faith has always played a crucial role in American political rhetoric (see, for example, Chilton 2004), and in contrast to Domke and Coe's (2010) strong claim of its dramatic rise in public speeches over the last two decades of the 20th century, this paper highlights the pervasive role of religion in American political culture. The second aim is methodological and lies in the complex interaction between quantitative and qualitative dimensions of diachronic analysis of any socio-political issue within specialized domains, as well as in the type of challenges facing the discourse analyst making use of corpora during the process. Finally, the paper will briefly tackle the no less important matter of how (diachronic) assisted discourse analysis can contribute to the study of history and politics in institutional domains.

2 The Corpus

The State of the Union (henceforth SoU) address is a significant public ritual, just like the inauguraladdress, the acceptance speech and other types of presidential rhetorical discourse (Campbell and Jamison 2008). In particular, the SoU is a constitutionally mandate form of address (Article II,

Section 3 of the U.S. Constitution), which is delivered, either written or oral by the President on a yearly basis. The SoU address is thus characterized by a specific discourse situation. Although it is primarily aimed at a very specific addressee, *i.e.* the Congress, over time, and especially since the advent of what scholars have defined the 'modern presidency' (see, for example, Tulis 1987), and thanks to its mediatisation, presidents have increasingly engaged with the American people at large.

The purpose of the address is chiefly to provide a general presentation of the President's agenda for the coming year, concerning both domestic and international political, social and economic priorities; moreover, aspects of legislative proposals are also included.

The SoU corpus includes all the 228 complete presidential transcribed speeches delivered by the 44 U.S Presidents since in January 1790 with President Washington's first address (which at the time was called the "Annual Message"), to President Obama's latest, made in January 2014. The corpus is 'clean' (*i.e.* not annotated) and amounts to about 1,800,000 running words. For the purpose of analysis, the corpus has been divided into five segments corresponding to main historical cleavages, as illustrated in Table 1 (see Bayley and Bevitori 2014, 2015; Bevitori 2014). However, the corpus can also be searched according to a range of different criteria; *i.e.* by president, by terms in office, by year(s), by party affiliation, or by a combination of any of them.

Segments	**1**	**2**	**3**	**4**	**5**
Years	1790-1864	1865-1916	1917-1945	1946-1989	1990-2014
Epoch	Up to last Civil War address	Before WW I	Up to end of WW II	Cold War	End of Cold War to present
Presidents	Washington to Lincoln	Johnson A. to Wilson	Wilson to Roosevelt F. D.	Truman to Bush G. H.	Bush G. H. to Obama
No. Addresses	76	52	28	47	25
No. Tokens	550,791	647,817	152,566	292,878	152,089

Table 1. Breakdown of the SoU corpus across historical periods

3 Methods and tools

In order to set out the goals and objectives that will frame the analysis, the tools and techniques of corpus linguistics and discourse analysis are used, and a corpus-assisted discourse analysis approach is proposed. This approach, as noted in a number of studies, (inter alia, Partington, Morley and Haarman (eds.) 2003; Baker 2006, 2011; Baker et al. 2008, Morley and Bayley (eds.) 2009), entails not only a blend of quantitative and qualitative dimensions in the analysis, but also, and perhaps more importantly, encourages the use of different research procedures in order to identify patterns of meaning across texts and contexts. However, while the approach has certainly proved fruitful, a number of issues have also been recently raised as regards its limits and constraints (see, for example, Miller et al. 2014, Bevitori 2015). As far as tools are concerned, Mike Scott's *WordSmith Tools 4.0* (Scott 2005) and *AntConc 3.2.4w* (Anthony 2011) were used for analysis.

4 Case study: 'may God bless'

Due to space (and time) constraints, the present study will set out to explore how and to what extent American presidents have appealed to God in their speeches. Previous comparative diachronic corpus-assisted analysis of interpersonal resources within this domain, *i.e.* the modal *may* (Bayley and Bevitori 2014), has in fact revealed that the modal, in the most recent historical period (segment 5), is most typically associated with 'God' and 'bless', thus making the formulaic 'may God bless' the most frequent three-word cluster in this segment. At close inspection, the cluster 'may God bless' across the whole corpus indicates that the use of this pattern began to emerge with president Truman, appearing for the first time, in the closing of his 1953 SoU address. Since then, variants of the three-word cluster have been, albeit sparingly, used by some of his successors (Kennedy 1962, Nixon 1970 and Ford 1977).

Nonetheless, it has been since the early 1980s, with President Reagan, that the phrase has been more regularly used in presidential addresses. According to Domke and Coe (2010), this tendency to make a deliberate and consistent use of faith in public speeches by U.S. presidents, marks the birth of a new "religious politics", as well as a powerful political tool or, as their title suggests, a 'God strategy'. This may certainly be confirmed by looking at patterns of the item *God* across all Reagan's SoU addresses, revealing that its occurrence accounts for almost half of all instances (49 percent) in the Cold war period (segment 4). Besides, a quantitative analysis of *God* across the whole corpus (Figure 1) reveals a dramatic and progressive increase in the frequency of occurrence of the word, coinciding with what is generally considered the beginning of the 'modern presidency' (see, for example, Tulis 1987), and in particular, since President FD Roosevelt in the mid-1930s. In fact, out of 157 occurrences, only 36 are found in the

period between 1790 and the early 1930s (23 and 77 percent respectively).

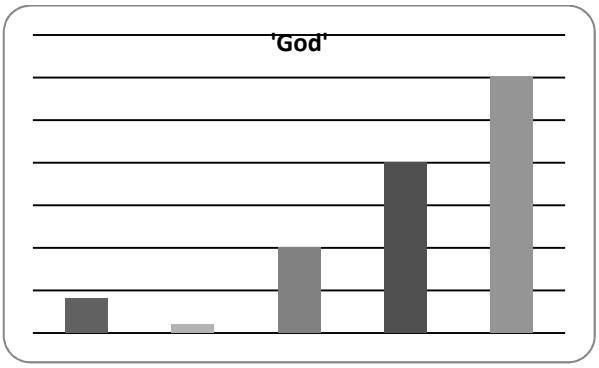


Figure 1. Relative frequency of *God* (per hundred tokens) across historical segments

Still, searching for 'God' is only but one part of the story. Word meanings may not be (and, indeed frequently, are not) stable over time and this, I believe, represents a great challenge to the corpus analyst attempting to combine quantitative and qualitative investigation of distinctive rhetorical structures over time (see also Bevitori 2015). In fact, close reading of most of the 18th and 19th century SoU addresses shows that there are numerous variants of the name 'God', which are difficult to retrieve only through the aid of concordances (Bayley and Bevitori 2014).

However, looking at the texts first can point to possible search terms which can be further explored through the software (e.g. *Providence, Supreme, Being, Divine Blessing,* etc). In particular, the analysis of the lemma *bless** across the same historical segments affords a somewhat useful complementary perspective. There are 271 occurrences of *bless** in the whole corpus (Figure 2), corresponding to 0.015 per hundred tokens.

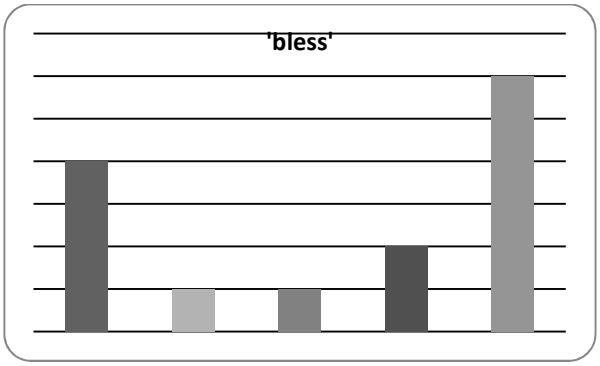


Figure 2. Relative frequency of *bless** (per hundred tokens) across historical segments

A comparison of segment 1 and segment 5 shows that while in the former the noun *blessing(s)* has a significantly higher frequency than the verb (79 percent of all instances (mostly in its plural form) compared to 21 percent of the latter), in the latter (segment 5), this tendency is completely the opposite. The verb, in fact, covers more than 80 percent of all instances. Moreover, in segment 1, patterns of 'blessing', in contrast to 'God', are frequently positioned at the beginning of the speech, invoking 'peace', 'health' and 'freedom'. These preliminary data will, however, need to be more thoroughly analysed in their wider co-text and context of occurrence to better explore differences and similarities across the different historical periods.

5 Conclusion

The study provides a necessarily limited and far from comprehensive view on the complex relationship between religion and politics in presidential discourse. Nonetheless, the paper offers possible routes to investigate patterns of change and stability across specialized diachronic corpora at the intersection between socio-political issues, methods and approaches.

References

Anthony, L. 2011. AntConc (Version 3.2.4w) [Computer Software] Tokyo, Japan: Waseda University. Available online at http://www.laurenceanthony.net/

Baker, P. 2006. *Using* Corpora *in* Discourse Analysis. London: Continuum.

Baker, P. 2011. "Times may change but we'll always have money: a corpus driven examination of vocabulary change in four diachronic corpora." In *Journal of English Linguistics*, 39: 65-88.

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyzanowski, M., McEnery, T. and Wodak, R. 2008, 'A useful synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press', in *Discourse and Society,* 19(3): 273-306.

Bayley, P. and Bevitori, C. 2011. "Addressing the Congress: Language change from Washington to Obama (1790-2011)". Unpublished Paper given at Clavier 11 International Conference , Tracking Language Change in Specialised and Professional Genres, University of Modena and Reggio Emilia, Modena, 24-26 November 2011.

Bayley, P. and Bevitori C. 2014. "In search for meaning: what corpora can/cannot tell. A diachronic case study of the State of the Union Addresses (1790-2013)". In Miller, D. R., Bayley, P., Bevitori, C., Fusari, S. and Luporini, A"'Ticklish trawling': The limits of corpus assisted meaning analysis'. In Alsop, S. and Gardner, S. (eds). Proceedings of ESFLCW 2013. Language in a Digital Age: Be Not Afraid of Digitality. 01-03 July 2013. Coventry University: Coventry, UK

Bayley, P. and Bevitori C. , 2015. "Two centuries of 'security': Semantic variation in the State of the Union

Address (1790-2014)". In Duguid, A., Marchi, A., Partington, A. and Taylor, C. (eds), *Gentle Obsessions: Literature, Language and Learning. In Honour of John Morley*. Roma: Artemide Edizioni.

Bevitori, C. 2014. "in a world of complex threats...": Discourses of (in)security in the State of the Union Address. A diachronic corpus-assisted study." Paper presented at the International Conference Critical Approaches to Discourse Analysis (CADAAD 5), Eötvos Loránd University, Budapest, 1-3 September 2014.

Bevitori, C. 2015 "Discursive constructions of the Environment in American Presidential Speeches 1960-2013: A diachronic corpus-assisted study". Baker, P. and McEnery, T. (eds) (2015) *Corpora and Discourse*. London: Palgrave.

Campbell, K. K., and Jamieson, K. H. 2008 *Presidents creating the presidency. Deeds done in word*s. Chicago: University of Chicago Press.

Chilton, P. 2004. *Analysing Political Discourse: Theory and Practice*. New York: Routledge.

Domke D. And Coe K. [2008] 2010. *The God strategy: How religion became a political weapon in America*. Oxford: Oxford University Press

Miller, D. R., Bayley, P., Bevitori, C., Fusari, S. and Luporini, A, "'Ticklish trawling': The limits of corpus assisted meaning analysis'. In Alsop, S. and Gardner, S. (eds). Proceedings of ESFLCW 2013. Language in a Digital Age: Be Not Afraid of Digitality. 01-03 July 2013. Coventry University: Coventry, UK.].Available online at http://curve.coventry.ac.uk/open/items/7b5b94aa-6984-48ad-b29a-9a8e9483fa2d/1/ ISBN: 978 18460007 13. [Full paper to appear in a collection of selected papers, Equinox 2015]

Morley, J. and P. Bayley (eds) 2009. *Corpus-assisted discourse studies on the Iraq conflict: Wording the war*. New York: Routledge

Partington, A., Morley, J. and Haarrman, L. (eds) 2004, *Corpora and Discourse*. Bern: Peter Lang.

Tulis, J. 1987. *The Rhetorical Presidency*. Princeton, NJ Princeton University Press.

Tagging and searching the bilingual public notices from 19th century Luxembourg

Rahel Beyer
University
of Luxembourg
rahel.beyer@uni.lu

1 Introduction

The project on the study of the language standardization of German in Luxembourg in the 19th century takes both macro-linguistic and micro-linguistic aspects into account in order to determine the effects of language policy on language practice and vice versa in 19th century Luxembourg. Accordingly, structural processes of variation and replication from contact languages as well as language policies and the different elements of language ideologies behind the language policies are analysed. Furthermore, the project draws upon several sources, some of which build a corpus of several thousand documents and of several million word forms as well as of mainly bilingual (French-German) texts.[25] In order to manage this amount of data, meet current quality standards in corpus linguistics and handle the special data type as well as project goals, an appropriate software was developed.

2 Structure and functions of the tool

The application should not only be a simple database, i.e. a data storage where you can filter for different subsets. Rather, it was intended to also have the possibility to distribute tags in the individual documents, search within the texts, show the results with context and process the findings.

As a consequence the php-application is divided into several parts. First of all, on the home page all data sets are listed. By entering characteristics in provided data fields the respective documents can be filtered. Several selection criteria can be combined to get to a specific subset.

By clicking on the signature the respective document opens in a new tab. Here, you can find the original scan of the public notice, the metadata (like signature, languages used, title, creation year etc.) and the full text. Values of the metadata can be revised or new metadata can be added. This metadata feeds the columns on the home page. Besides metadata tagging every token of the full text can be tagged for various information and for

[25] For text selection and compilation see Gilles/Ziegler (2013), for further information on the project see Beyer et al. (2014).

various aspects. So, whenever there is a variant of a linguistic phenomenon in question, a tag documenting the instance can be assigned. The same works for expressions or even phrases referring to a certain language attitude. There are no predefined or mandatory tags, instead, they can individually and flexibly created. Although in the provided window the last allocated tag (i.e. tag name and value) is specified and can be used directly. Furthermore, a search request for a selected token can be generated out of the current document, i.e. a tab with a search form prepopulated for item (the selected token), tags and language opens. Also, by mouseover on the tokens links to different dictionaries can be followed.

Given the bilingual edition of the texts, separation of the two languages involved was a crucial requirement. The two language versions of the text can be displayed either one below the other in one column or in separated columns next to each other. The language were assigned to the paragraphs automatically on the basis of different font types. There is also the possibility to change the language assignment in the documents.

In the search menu you can also specify a context word and its maximum distance to the searched expression. Via specification of metadata information you can further narrow down the relevant documents which should be searched.

The display of search results corresponds to those of typical concordance programs, but again it's more flexible, i.e. you can choose other, respectively more metadata to be shown than only the signature of the documents. Additionally, in the right column you have the possibility to deactivate results, so that they won't be proceeded any further. In case all (remaining) findings are supposed to get the same tag and respective value, the option "apply tag to all search results" can be chosen.

In the statistics menu all tags are listed and counted. The sums are presented in a table. This gives you the quantitative analysis and provides information on the distribution of variants in the course of the period of analysis. The division in subperiods can be adapted to individual needs. From the statistics you can also get to the instances.

3 Conclusion

In sum, a tool was developed being flexible to a great extent. The display of information can individually selected, tags of personal choice can be created and settings can be adapted according to requirements.

References

Beyer, R., Gilles, P., Moliner, O. and Ziegler, E.. 2014. "Sprachstandardisierung unter Mehrsprachigkeitsbedingungen: Das Deutsche in Luxemburg im 19. Jahrhundert". *Jahrbuch für germanistische Sprachgeschichte* 5: 283-298

Gilles, P. and Ziegler, E.. 2013. "The Historical Luxembourgish Bilingual Affichen Database"". In P. Bennett, M. Durrell, S. Scheible and R.J. Whitt (eds.) *New methods in Historical Corpus Linguistics*. Tübingen: Narr. 127-138

A linguistic taxonomy of registers on the searchable web: Distribution, linguistic descriptions, and automatic register identification (*panel*)

Doug Biber
Northern Arizona University
Douglas.Biber@nau.edu

Jesse Egbert
Brigham Young University
Jesse_Egbert@byu.edu

Mark Davies
Brigham Young University
Mark_Davies@byu. edu

1 Introduction

For both general users and linguists, the advantages of the World Wide Web are obvious: it provides a massive amount of linguistic data, readily accessible to anyone with a computer. However, the nature of the different types of language used on the web remains unclear. In particular, we currently have little information about the text categories—the 'registers'—found on the web.

The mystifying composition of the Web is especially problematic for linguists using the web as a corpus to investigate linguistic patterns of use. This approach has become so prevalent that the acronym WAC (Web-as-Corpus) is now commonplace among researchers who explore ways to mine the WWW for linguistic analysis. One of the major challenges for WAC research is that a typical Web search usually provides us with no information about the kinds of texts investigated (see Kilgarriff and Grefenstette 2003).

These concerns are shared widely among WAC researchers, and as a result, there has been a surge of interest over the last several years in Automatic Genre Identification (AGI): computational methods using a wide range of descriptors to automatically classify web texts into genre—or register—categories. Of course, the prerequisite for computational techniques that automatically identify the register of a web document is a taxonomy of the possible register categories found on the web. That is, it is not possible to develop and test methods for the automatic prediction of register until we know the full set of possible web registers. In addition, it would be beneficial to know the distribution of those registers: which ones are especially prevalent, and which ones are rare. To date, however, efforts to obtain this information have had limited success.

We present the results of three research studies, each building on the results of the preceding one, leading to the goal of Automatic Register (or Genre) Identification of web documents. In the first study, we developed methods for user-based identification of the register category of web documents, and applied those methods to a corpus of 48,571 web documents. Based on the results of this first study, we are able to document the types and distribution of registers found on the searchable web, including the prevalence of 'hybrid' registers. In the second study, we carried out comprehensive lexico-grammatical analyses of each web document in our corpus, leading to a Multi-Dimensional (MD) analysis of the patterns of register variation found on the web. Finally, in the third study, we evaluated the predictive power of our MD analysis, analyzing the extent to which these linguistic characteristics can predict the register category of new web documents.

2 Corpus for analysis

The corpus used for the study was extracted from the 'General' component of the Corpus of Global Web-based English (GloWbE; see http://corpus2.byu.edu/glowbe/). The GloWbE corpus contains c. 1.9 billion words in 1.8 million web documents, collected in November-December 2012 by using the results of Google searches of highly frequent English 3-grams (i.e., the most common 3-grams occurring in *COCA*; e.g., *is not the, and from the*). 800-1000 links were saved for each n-gram (i.e., 80-100 Google results pages), minimizing the bias from the preferences built into Google searches. Many previous web-as-corpus studies have used similar methods with n-grams as search engine seeds (see, e.g., Baroni & Bernardini, 2004; Baroni et al., 2009; Sharoff, 2005; 2006). It is important to acknowledge that no Google search is truly random. Thus, even searches on 3-grams consisting of function words (e.g., *is not the*) will to some extent be processed based on choices and predictions built into the Google search engine. However, selecting hundreds of documents for each of these n-grams that consist of function words rather than content words minimizes that influence.

To create a representative sample of web pages to be analyzed in our project, we randomly extracted 53,424 URLs from the GloWbE Corpus. This sample, comprising web pages from five geographic regions (United States, United Kingdom, Canada, Australia, and New Zealand), represents a large sample of web documents collected from the full spectrum of the searchable Web. Because the ultimate objective of our project is to describe the lexico-grammatical characteristics of web documents, any page with less than 75 words of text was excluded from this sample.

To create the actual corpus of documents used for our study, we downloaded the web documents associated with those URLs using HTTrack

(http://www.httrack.com). However, because there was a 7-month gap between the initial identification of URLs and the actual downloading of documents, c. 8% of the documents (n = 3,713) were no longer available (i.e., they were linked to websites that no longer existed). This high attrition rate reflects the extremely dynamic nature of the universe of texts on the Web.

Our ultimate goal in the project is to carry out linguistic analyses of internet texts from the range of web registers (see discussion in the conclusion). For this reason, 1,140 URLs were excluded from subsequent analysis because they consisted mostly of photos or graphics. Thus, the final corpus for our project contained 48,571 documents. To prepare the corpus for POS tagging and linguistic analyses, non-textual material was removed from all web pages (HTML scrubbing and boilerplate removal) using JusText (http://code.google.com/p/justext).

3 Study 1: A user-based taxonomy of web registers

For the first study, we employed a bottom-up user-based investigation of a large, representative corpus of web documents. Instead of relying on individual expert coders, we recruit typical end-users of the Web for our register coding, with each document in the corpus coded by four different raters. End-users identify basic situational characteristics of each web document, coded in a hierarchical manner. Those situational characteristics lead to general register categories, which eventually lead to lists of specific sub-registers. By working through a hierarchical decision tree, users are able to identify the register category of most internet texts with a high degree of reliability.

The approach we have adopted here makes it possible to document the register composition of the searchable web. Narrative registers are found to be the most prevalent, while Opinion and Informational Description/Explanation registers are also found to be extremely common. One of the major innovations of the approach adopted here is that it permits an empirical identification of 'hybrid' documents, which integrate characteristics from multiple general register categories (e.g., opinionated-narrative). These patterns are described and illustrated through sample internet documents.

4 Study 2: Comprehensive lexico-grammatical description of web registers

Study 2 begins by considering the patterns of register variation with respect to the Biber (1988) linguistic dimensions. These analyses show that there are major linguistic differences among the eight major user-defined register categories. For example Figure 1 plots the Dimension 1 scores for these web registers (shown in **BOLD CAPS**) compared to five major non-internet registers.

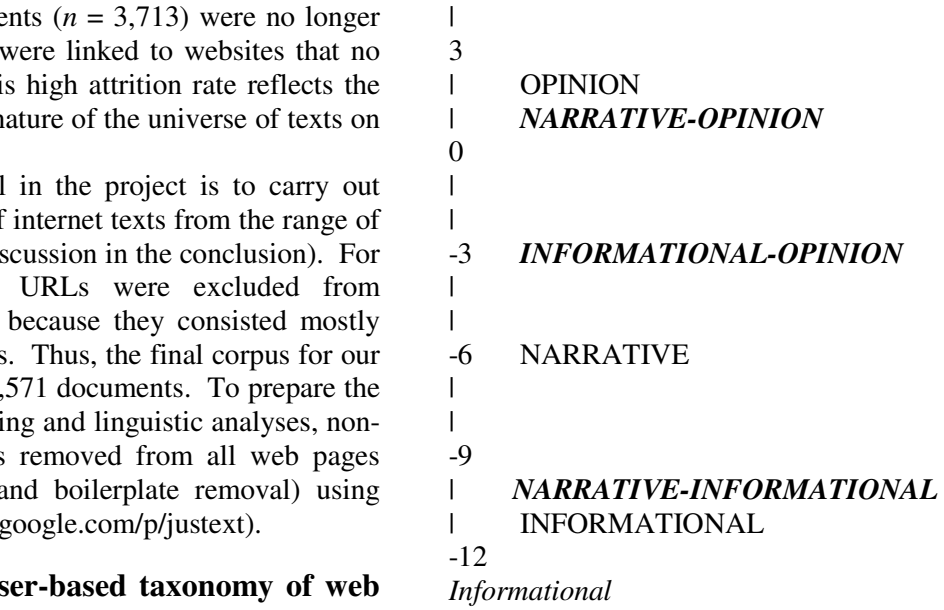


Figure 1. Dimension 1 Scores for the 8 Online Registers and 5 Registers from Biber (1988).

Interestingly, the 'hybrid' registers identified by end-users behave in hybrid ways with respect to the 1988 dimensions. For example, Figure 2 plots the scores for three hybrid registers (shown in BOLD ITALICS) along Dimension 1, showing how they have consistently intermediate scores between the associated simple registers.

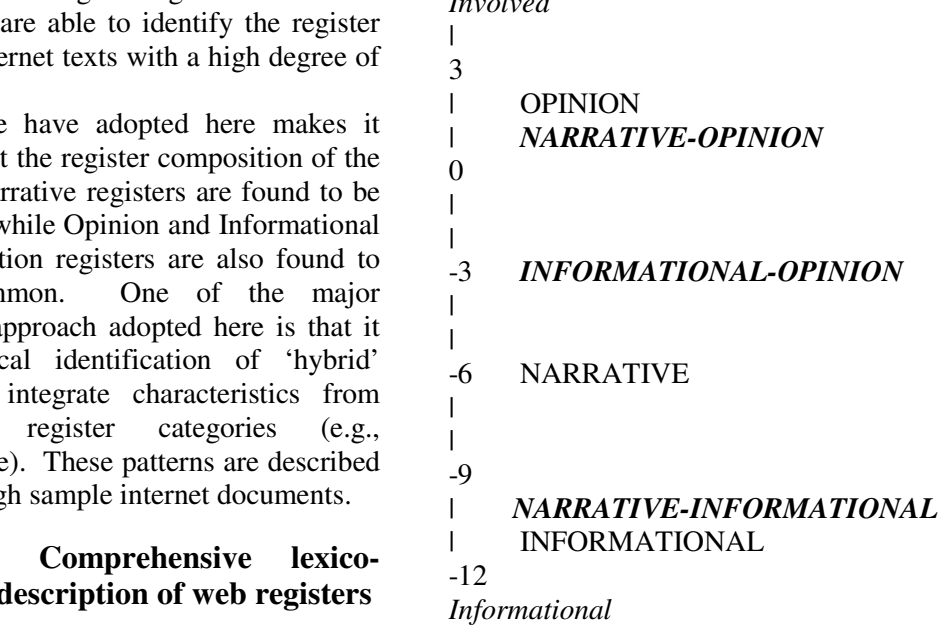


Figure 2. Dimension 1 Scores for 3 Simple Registers and 3 Hybrid Registers.

Building on these findings, we carried out a new factor analysis to identify the linguistic dimensions of variation that are well-defined in this discourse

domain. The primary focus of Study 2 is on the linguistic composition of those dimensions, and the patterns of register variation along each one.

5 Study 3: Evaluation of the linguistic description: Automatic Register Identification

Finally, in Study 3 we carried out an evaluation of the linguistic description, describing the extent to which these linguistic variables can accurately predict the register categories of web documents. For this purpose, we reserved a random sample of c. 10,000 web documents that had been coded for register characteristics in Study 1, but not used for the MD analysis in Study 2. Thus, we are able to directly evaluate the extent to which the linguistic dimensions of variation identified in Study 2 can correctly determine the register category of 'new' web documents.

References

Baroni, M and Bernardini, S. 2004. BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004, Lisbon: ELDA. 1313-1316.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43 (3): 209-226.

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Kilgarriff, A. and Grefenstette, G. 2003. Introduction to the special issue on the Web as corpus. *Computational Linguistics*, 29:333-347.

Sharoff, S. 2005. Creating general-purpose corpora using automated search engine queries. In M. Baroni and S. Bernardini, (Eds.), *WaCky! Working papers on the Web as Corpus.* Gedit, Bologna.

Sharoff, S. 2006. Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics,* 11(4), 435-462.

On the (non)utility of Juilland's D for corpus-based vocabulary lists

Douglas Biber
Northern Arizona University
Douglas.biber@nau.edu

Randi Reppen
Northern Arizona University
Randi.reppen@nau.edu

Erin Schnur
Northern Arizona University
Erin.Schnur@nau.edu

Romy Ghanem
Northern Arizona University
Rg634@nau.edu

Vocabulary lists of the most important words in a discourse domain have become increasingly popular in recent years, and corpus linguists have been at the forefront of the efforts to develop lists that have high validity (e.g., word lists representing general English, or word lists representing academic writing; see, e.g., Coxhead 2000, Leech et al. 2001, Davies and Gardner 2010, Gardner and Davies 2013, Brezina and Gablasova 2013). Two major factors are considered when carrying out the corpus analyses to develop such lists: frequency and dispersion. It is universally accepted that frequency is a required criterion for identifying the important words in a corpus. But an equally important consideration is dispersion: a word that is both frequent and widely distributed across an entire corpus is more 'important' than a high-frequency word that is restricted to one or two texts. As a result, nearly all recent corpus-based vocabulary researchers consider both frequency and dispersion measures when constructing new vocabulary lists.

One measure of dispersion – Juilland's D – is usually used in these projects, because it is widely regarded as the most reliable measure (see, e.g., Lyne 1985, Leech et al. 2001, Davies and Gardner 2010). However, Gries (2008) raises serious criticisms of this measure (as well as most other previous measures of lexical dispersion in corpora), proposing instead an alternative measure: DP ('Deviation of Proportions').

In theory, Juilland's D is claimed to have a range of 0.0 to 1.0, with values close to 1.0 representing a completely even dispersion of a word in a corpus. However, in our own recent research, we began to note that the reported D values for specific words were often at odds with our perceptions of how specialized a word was. For example, in the Leech et al (2001) general word list (based on both the spoken and written components of the BNC, divided into 100 corpus parts), many words that we associated with academic writing (e.g., *however,*

thus, presence, political, illustrate, and *implement*) had D values over .90, indicating a uniform distribution across the entire BNC. Similarly, in the Davies and Gardner general word list (based on the spoken and written components of COCA, divided into 388 corpus parts), specialized words like *random, guilt, strain, behave, crystal, execute, motive, convict,* and *simultaneously* all had D values over .90. In contrast, it was extremely difficult to locate any words with D values lower than .50 in either of these word lists. (For example, even *ooh* in the BNC word list has a D value of .58, and *ok* in the COCA word list has a D value of .78.)

Closer inspection of the formula for Juilland's D reveals that the measure is directly dependent on the number of corpus parts used for the analysis: as the number of parts becomes large, the value for D approaches 1.0. Early uses of D in the 1950s and 1960s were based on corpora divided into a few parts, and thus this characteristic of the measure was not problematic. In more recent applications, though, corpora have been divided into 100-500 parts, minimizing the effective range of the scale for D so that it becomes much less useful as an indicator of dispersion.

We illustrate this mathematical relationship through a series of experiments based on analysis of the BNC. Two variables are manipulated in our experiments: the distributional characteristics of the words being analyzed, and the number of corpus parts used for the analysis. We analyzed a sample of 185 different words, chosen to represent the range of possible distributional profiles that words might have in the BNC. The first set of words were selected from the 'distinctiveness list contrasting speech and writing', presented as Table 2.4 in Leech et al. (2001). Our goal with this set of words was to include words that are clearly skewed in distribution, with the prediction that those words should have low D values. For this sample, we chose words that had the largest Log Likelihood values. (Half of the words in our sample had large positive LL scores, reflecting their skewed use in speech, and half of the words in our sample had large negative LL scores, reflecting their skewed use in writing.) We further sampled words from four frequency bands, to analyze the influence of frequency on the D measure. As a result, we considered samples of words from eight different categories:

- Frequency ≥ 4,000 per million words; LL distinctive for Speech (e.g., *er, you, we, I, yeah*)
- Frequency 3,999 – 500 per million words; LL distinctive for Speech (e.g., *mm, think, cos, yes, put*)
- Frequency 499 – 101 per million words; LL distinctive for Speech (e.g., *ooh, eh, hello, aye, aha*)
- Frequency ≤ 100 per million words; LL distinctive for Speech (e.g., *bloke, ha, bet, urgh, reckon*)
- Frequency ≥ 4,000 per million words; LL distinctive for Writing (e.g., *the, in, from, with, had*)
- Frequency 3,999 – 500 per million words; LL distinctive for Writing (e.g., *an, however, may, between*)
- Frequency 499 – 101 per million words; LL distinctive for Writing (e.g., *thus, social, began, among*)
- Frequency ≤ 100 per million words; LL distinctive for Writing (e.g., *smiled, latter, Fig., Inc, methods*)

In addition, to represent words that likely have even distribution across the BNC, we selected a sample of 40 words with extremely high values for D (>= .97) in the Leech et al. (2001) 'alphabetical frequency list' (Table 1.1). We grouped these words into two major frequency bands:

- Frequency > 500 per million words; D >= .97 (e.g., *all, and, as, before, but*)
- Frequency ≤ 500 per million words; D >= .97 (e.g., *able, bring, brought, decided*)

The primary focus of the study was to determine the influence of N – the Number of corpus parts – on the value for Juilland's D. For this purpose, we initially divided the BNC into 1,000 equal-sized parts (with each part containing 100,000 words), and computed the D value for each of the words in our sample. We then combined adjacent parts, to manipulate the value of N in the formula for D, carrying out separate series of computations for N = 500, 100, 50, 20, 10, and 5.

It is anticipated that the results of the experiments will demonstrate the strong influence of N – the number of corpus parts – on the effective scale for Juilland's D. We predict that experiments with high values for N will have a restricted range of values for D, while experiments with lower values for N will display a much greater range of variation. For comparison, we compute values for Gries' DP dispersion measure, which is predicted to consistently have an effective range of 0 – 1.0, regardless of the number of corpus parts.

References

Brezina, V. and D. Gablasova. 2013. Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*. 1-23

Coxhead, Averil. 2000. A new academic word list.

TESOL Quarterly 34: 213-238.

Davies, M. and D. Gardner. 2010. *A Frequency Dictionary of Contemporary American English: Word Sketches, Collocates, and Thematic Lists*. Routledge.

Gardner, D. and M. Davies. 2013. A new academic vocabulary list. *Applied Linguistics*, 34(5), 1-24.

Leech, G., P. Rayson, and A. Wilson. 2001. *Word frequencies in written and spoken English: Based on the British National Corpus*. Longman.

Licensing embedded sentence fragments

Felix Bildhauer
Freie Universität
felix.bildhauer@fu-berlin.de

Arne Zeschel
Institut für Deutsche Sprache
zeschel@ids-mannheim.de

1 Introduction

In situated interaction, the propositional argument of several German complement-taking predicates (CTPs) may be realised in three different ways: (i) as a canonical subordinate clause (involving a complementiser and verb-final word order, cf. 1.a), (ii) as an apparent main clause (no complementiser and verb-second/v2 word order, cf. 1.b) and (iii) as an elliptical focus construction in which individual constituents of various grammatical functions can stand in for a complete clause (cf. 1.c):

(1) A: Kommen sie heute oder morgen?
'Will they come today or tomorrow?'

B: a. Ich denke, dass sie morgen kommen.
b. Ich denke, sie kommen morgen.
c. Ich denke morgen.
'I think [(that) they will come] tomorrow'.

At the same time, even verbs with highly similar meanings do not behave alike in this respect:

(2) a. Ich denke, …
… dass sie morgen kommen.
… sie kommen morgen.
… morgen.

'I think [(that) they will come] tomorrow'.

b. Ich weiß, ...
… dass sie morgen kommen.
… sie kommen morgen.
… *morgen.

'I know [(that) they will come] tomorrow'.

c. Ich bezweifle, …
… dass sie morgen kommen.
… ?? sie kommen morgen.
… *morgen.
'I doubt [(that) they will come] tomorrow'.

Since they are not licensed across the board, frag-

mentary complement clauses like (1c) cannot be accounted for by unspecific appeals to 'recoverability in context' alone. How can the contrasts in (2a-c) be explained, then? We explore the possibility that types of permitted ellipses can be predicted from governing verbs' preference for particular kinds of non-elliptical complement clauses. For instance, wissen 'know' most commonly combines with wh-clauses among its sentential complements. And though ungrammatical in the ellipsis in (2.b), it works well with 'sluices' (Ross 1969) such as (3):

(3) Ich habe es schon mal gesehen, aber ich weiß nicht, wo.

'I have seen it before, but I don't know where'

On this account, acceptable ellipses like (1.c) and (3) would be conventionalised fragments of CTP's most entrenched schemas for 'canonical' sentential complementation (and only these), and there is no independent functional or semantic generalisation behind the contrasts in (2). On the other hand, also if permissible ellipses were indeed conventionalised fragments of a CTP's most common complementation pattern, they could still be subject to idiosyncratic further constraints: for instance, in contexts that license such ellipses in principle (e.g. question-answer adjacency pairs), it is conceivable that an ellipsis may nevertheless require additional pragmatic preconditions to be met that are irrelevant to its non-elliptical counterpart. In this case, any such additional constraints would need to be captured somewhere in the grammar.

2 Corpus study

We explore these issues in a combined corpus-linguistic and interactional study of 25 CTPs from different semantic classes using samples of 500 attestations each. The data is taken from the German national conversation corpus FOLK (Deppermann & Hartung 2011) and a subset of the DECOW2012 web corpus (Schäfer & Bildhauer 2012) containing 'quasi-spontaneous' CMC data.

Before coding the full set of 25x500=12500 samples, we conducted a pilot study with seven verbs from three semantic classes:

EPISTEMIC STATUS
denken 'to think'
wissen 'to know'
bezweifeln 'to doubt'

PROPOSITIONAL ATTITUDE
fürchten 'to fear'
befürchten 'to fear'

SOURCE OF KNOWLEDGE
hören 'to hear'
merken 'to notice'

For each of these verbs, we obtained the proportions of different types of complete (*dass* vs. v2 vs. wh) and fragmentary complement clauses (*dass-*/v2-substituting vs. wh-substituting ellipses) in our sample. Next, these results were subject to correlation analysis.

3 Results

For *dass*/v2-substituting ellipses, we found a strong correlation between the occurrence of fragmentary complement clauses and verbs' bias for v2-complement clauses ($r_{Pearson}$=.85, p=.01). No such correlation was found for verbs' bias for *dass*-clauses (r=-.33, p=.42) or the percentage of both kinds of complement clauses taken together (r=.52, p=.24). This suggests that the occurrence of non-wh-substituting ellipses cannot be predicted from verbs' biases for *dass*- or non-wh-sentential complementation in general. Rather, relevant expressions appear to be modelled on a v2-complementation schema. For wh-substituting ellipses, preliminary results could not yet be obtained since most verbs in the pilot study proved semantically incompatible with wh-complementation (yielding no hits for relevant clauses or ellipses at all).

4 Outlook

Since the results of the pilot study point in the expected direction, the study is currently expanded to the full set of 25 CTPs (12,500 data points). The expanded version comprises five different verbs from five semantic classes, including a greater number of wh-compatible types. In a first step, we repeat the procedure outlined above for the total dataset. Next, we zoom in on the actual usage patterns of the elliptical utterances thus identified by investigating a variety of their morphosyntactic, semantic, deictic, information structural and sequential context properties. We close with a brief discussion of theoretical options for modelling our findings within a surface-oriented, construction-based approach to grammar: what is the theoretical status of structures (i.e. our fragmentary complement clauses) that are apparent variants of other constructions (i.e. full syntactic complementation patterns), in particular if these structures are not merely different in form but also show a more restricted distribution?

References

Deppermann, A. and Hartung, M. 2011. "Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme

und Prioritäten der Stratifikation des 'Forschungs- und Lehrkorpus Gesprochenes Deutsch' (FOLK) am Institut für Deutsche Sprache (Mannheim)". In Felder, E., et al. (eds.) *Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen*. Berlin, New York: de Gruyter, 414-450.

Klein, W. 1993. "Ellipse". In J. Jacobs et al. (ed.), *Syntax. Vol. 1. Ein internationales Handbuch zeitgenössischer Forschung*. Berlin: de Gruyter, 763-799.

Ross, J. R. 1969. "Guess who?" In R. Binnick, et al. (eds.) Papers from the 5th regional meeting of the Chicago Linguistic Society. Chicago, Ill.: Chicago Linguistic Society, 252-286.

Schäfer, R. and Bildhauer, F. 2012. "Building large corpora from the web using a new efficient tool chain". In Nicoletta Calzolari, et al. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association, 486-493.

Forward-looking statements in CSR reports: a comparative analysis of reports in English, Italian and Chinese

Marina Bondi
University of Modena and Reggio Emilia
marina.bondi@unimore.it

Danni Yu
University of Modena and Reggio Emilia
dannimail@foxmail.com

With the current international awareness of environmental issues, corporate social responsibility (CSR) has attracted great attention from practitioners and researchers. The CSR report, a genre which has become standard practice in CSR disclosure, has been studied by linguists from various perspectives (Bondi forthcoming, Catenaccio 2012, Malavasi 2012, Fuoli 2012, Fuoliand Paradis 2014; Bhatia 2012, 2013; Wang 2013). While clearly reporting past action and performance in the field of CSR, the genre also presents elements of outlook and references to future action which contribute greatly to the construction of corporate identity. The frequency, scope and function of these references to the future still remain to be studied in depth.

The paper presents a corpus-based exploration of how forward-looking statements are realized in different languages.The corpus is composed of 90 CSR reports in English, Italian and Chinese in two main sectors: energy (electricity, oil or gas) and banking. These were chosen from the top-ranking institutions in Italy, China and worldwide. While the entire corpus is used to investigate general linguistic features of certain moves, a subcorpus of 18 reports has been analyzed with the aim to establish a move-step scheme which has later been adopted to annotate the same sub-corpus.

The analysis starts from an overview of the role of forward-looking statements within the structure of the report itself, looking at whether these forward-looking statements are spread all through the structure of the report or rather appear more frequently in a final outlook section.The data of the move 'Previewing future performance' in the small corpus analyzed shows that forward-looking statements are used most in CSR reports in English, and least in Chinese. The data also shows that futurity is more marked in the macro-move 'Performance-reporting' than in 'Self-presenting'.

The analysis then looks at particular linguistic devices that are used to express futurity.The notion of futurity is recalled in many ways in the conceptual structure of different lexical elements, in

the way they categorize reality (Bondi, forthcoming, p.10). The first phase of the analysis led to identifying future references of these lexical sets and their frequencies. Concordance analysis was then aimed at studying collocation, semantic preference and pragmatic function of the items in context.

The paper primarily looks at two sets of words and their derivatives: *improve/migliorare/改善* and *continue/continuare/ 继续*. Although their future meaning is not as obvious as that of words like *future* and *will*, these verbs are frequently used to indicate future action. They also both refer to the future by reference to the present, representing this connection as a process of change (*improve/migliorare/ 改善)* or as an element of continuity *(continue/continuare/继续).*

As for *improve/migliorare/改善*, we find that in English the frequency of *improve* is four times higher than *improved*, while in Italian *migliorare* is ten times as frequent as *migliorato*, and in Chinese *改善* is eight times more frequent than *改善了*. With a qualitative analysis looking at the concordances of the words *improve/migliorare/改善*, we rarely find examples that are used to indicate past meaning. Hence it could be inferred that in CSR reports verbs of change, such as *improve/migliorare/ 改善*, are more often used with future meaning. But we also notice an interesting difference between Italian and English. A comparison with reference corpora representative of the two languages taken into account (CORIS for Italian, COCA for English) has highlighted that in CORIS the token *migliorare* (4217) is almost eight times more frequent than *migliorato* (575), whereas in COCA, *improve* (28412) is not even one time more frequent than *improved* (17423). Reference to future improvement thus seems to be more frequent in CSR reports than in general corpora, especially in English.

In a similar way, the words *continue/continuare/ 继续* seem to be also more frequently used in forward-looking statements than with past reference in our corpora. The future use of *continue,* for example, is twice more frequent than its past use. Reference to elements of continuity is shown to contribute to the representation of the company's present identity and values.

Concordance analysis – in terms of collocates and semantic preference (Sinclair 2004) of the lexical items - also provides interesting data as to the areas of improvement and the types of processes to be continued most frequently found in the three corpora.

Finally, we look at how these elements are involved in expressing **prediction or commitment**. Phraseology is also studied in terms of semantic sequences (Hunston 2008) to highlight how prediction and commitment statements are used to construct corporate values.

In conclusion, the analysis brings to light some common communicative strategies used by different companies when projecting their future intentions while using forward-looking expressions, while differences in item frequency across corpora reveal particular linguistic preferences in different cultures.

References

Bhatia, A. 2012. "The CSR report-the hybridization of a confused genre (2007-2011) research article". *IEEETransactions on professional communication* 55(3): 221-228.

Bhatia, A. 2013. "International genre, local flavour-analysis of PetroChina's Corporate and Social Responsibility Report". In *RevistaSignos. Estudios de linguistica*. 46(83): 307-331.

Bondi, M. Forthcoming. "The future in reports: prediction, commitment and legitimization in CSR". *Pragmatics and society*.

Catenaccio, P. 2012. *Understanding CSR Discourse: Insights from Linguistics and Discourse Analysis*. Milano: Brossura.

Fuoli, M. 2012. "Assessing Social Responsibility: A quantitative analysis of Appraisal in BP's and IKEA's social reports." *Discourse & Communication* 6(1): 55-81.

Fuoli, M. andParadis, C. 2014. "A model of trust-repair discourse."*Journal of Pragmatics* 74: 52-69.

Hunston, S. 2008. "Starting with the small words: Patterns, lexis and semantic sequences". In *International Journal of Corpus Linguistic s* 13/3: 271-295.

Malavasi, D. 2012. 'The necessary balance between sustainability and economic success', an analysis of Fiat's and Toyota's Corporate Social Responsibility Reports." In P. Heynderickxet al (eds.) *The Language Factor in International Business*. Bern: Peter Lang, 247-264.

Sinclair, J.M. 2004. *Trust the Text. Language, Corpus and Discourse*. London: Routledge.

Wang, D. 2013. "Applying Corpus Linguistics in Discourse Analysis". In *Studies in Literature and Language* 6(2): 35-39.

Depictions of strikes as "battle" and "war" in articles in, and comments to, a South African online newspaper, with particular reference to the period following the Marikana massacre of August 2012

Richard Bowker
Rhodes University, Grahamstown
r.bowker@ru.ac.za

Sally Hunt
Rhodes University, Grahamstown
s.hunt@ru.ac.za

Even 21 years after the official end of apartheid, South Africa is a country of vast socio-economic inequalities, and, partly as a result of this, consistently experiences high numbers of labour strikes every year (especially during the so-called "strike season", the third quarter of the year). This was especially the case during the 2008-2012 period. One of these, the unprotected strike at Lonmin's Marikana platinum mine in August 2012, resulted in the massacre of 34 striking mineworkers by the South African police, in addition to the killings of 10 people in the previous week. Following Marikana, a series of unprotected strikes in 2012 spread to other locations and other mining sub-sectors in the country.

It is a well-established claim that language articulates and perpetuates ideologies (e.g. Fairclough 2001, 2010). It is also fairly well-established that the growing field of Computer-Assisted Discourse Studies – such as the analysis that can be undertaken using a combination of Corpus Linguistics and Critical Discourse Analysis – can provide a means of accessing the ideologies that inhere in discourse. Beginning with the observation that strikes and related events and processes are routinely depicted as "battle" and "war" – for example, in photo gallery links like "Photos from the front lines" (Letsoalo 2011), and in headlines such as "Pay wars sideline job creation" (Donnelly 2011) and "Battle lines drawn as municipal wage negotiations begin" (SAPA 2012) – we set out to investigate the mechanisms whereby strikes are depicted in this manner, the discourses that are operationalised in such representations (Fairclough 2010), and the ideologies that are activated in the process.

In order to do so, we constructed a specialised corpus of online newspaper articles in the national weekly *Mail & Guardian* and the online comments to them, covering the period January 2008 to December 2012. The corpus comprises the majority, if not all accessible, online articles and comments in that newspaper over that the period that had strikes in South Africa as their topic, together making up just over 1.03-million tokens, of which slightly more than half are from the newspaper articles. Spurred on by the national debate that followed the massacre, nearly half of the tokens (roughly equally from articles and comments) are from 2012, and most of these are from the second half of 2012.

In order to understand what was happening in these representations, we examined in some detail the concordances showing "battle" and "war". It became clear that while most of the uses of "battle" and around half of the uses of "war" were metaphorical constructions (in the manner described by Lakoff & Johnson 1980), that could be classified according to the categories Political & Ideological, Labour & Class, and Strikes, there were also a fair number that were not metaphorical at all, but were literal depictions of strikes as "battle" or "war". That is, while the event that took place was a strike, or a series of strikes, the semiotic object (Peirce 1955) constructed was a battle or a war – in other words, it is not that the strike is *like* war; the strike *is* war. The question then arises: *how are "battle" and "war" adequate representations of strikes?*

Focusing on these literal depictions of strikes as "battle" and "war" (but excluding those that spoke of actual historical or distal contemporary wars), we found that many of them – particularly in the online comments sections, and especially after Marikana – indicated that, with regard to the strike events, a war, specifically a "civil war", was seen as being underway in South Africa or was something to which the country was heading. This is quite clearly not the case in reality, yet the commenters found it convenient to depict the series of strikes in this manner, or may actually have believed this to be the case.

In the main analytical section of this paper we examine in detail the concordances of the comments that express this point of view, and investigate what other representations of strikes and strikers and strikers are bundled together with these descriptions. We find that commenters tend to resort to assumptions and generalisations (Machin & Mayr 2012) in the form of illicit fusions and illicit fissions (Bhaskar 1993) with little or no epistemological warrant in order to force the strikers into particular negatively valenced representations. We also find that such representations increase well above the average that might be expected during periods of high incidence of strikes and of high volumes of reports on strikes. This is a matter of the construction of new meanings on the part of the commenters based on their interpretation of the events reported in the articles, in which a dominant discourse was fairly quickly established, and on their own ideological leanings.

We consider the event of the Marikana massacre (and related processes at the time) an outbreak of coercion on the part of state executive forces resulting from the beginnings of a breakdown in the hegemony of the Tripartite Alliance [26] over the working class (Gramsci 1971). We conclude that representing strikes as battles or wars allows the commenters to depict the strikers as an undifferentiated bloc (a war party) that is then separated out from conventional notions of South African society and nation; that is, striking workers are not seen as members of South African society. These representations mean that the strikers are seen as enemy combatants – people against whom a civil war might be fought. In the main, these commenters regress to colonial discourses or stagnate in blinkered middle-class perspectives (Fanon 1963) with such representations of the strikes. Finally, we discuss what such representations might mean with regard to the middle-class commenters' own positioning within South African society.

References

Bhaskar, R. (1993) *Dialectic: The Pulse of Freedom*. London: Verso

Donnelly, L. (2011) "Pay wars sideline job creation" in *Mail & Guardian*, 22 July, http://mg.co.za/article/2011-07-22-pay-wars-sideline-job-creation (Accessed 06.08.13)

Fairclough, N. (2001) *Language and Power*. 2nd edition. Abingdon: Routledge

Fairclough, N. (2010) *Critical Discourse Analysis: The Critical Study of Language*. 2nd ed. Harlow: Longman

Fanon, F. (1963) *The Wretched of the Earth*. Trans. Farrington, C. New York: Grove Press

Gramsci, A. (1971) *Selections from the Prison Notebooks*. Trans. Hoare, Q. & Nowell-Smith, G. London: Lawrence & Wishart

Lakoff, G. & Johnson, M. (1980) *Metaphors we live by*. Chicago: University of Chicago Press

Letsoalo, M. (2011) "SA hit by strike fever" in *Mail & Guardian*, 15 July, http://mg.co.za/article/2011-07-15-sa-hit-by-strike-fever (Accessed 06.08.13)

Machin, D. & Mayr, A. (2012) *How to Do Critical Discourse Analysis*. London: Sage

Mail & Guardian www.mg.co.za

Peirce, C. S. (1955) *Philosophical Writing of Peirce*. Ed. Buchler, J. New York: Dover

SAPA (2012) "Battle lines drawn as municipal wage negotiations begin" in *Mail & Guardian*, 22 May, http://mg.co.za/article/2012-05-22-municipal-wage-talks-begin (Accessed 07.08.13)

[26] The Tripartite Alliance consists of the ruling African National Congress, the South African Communist Party, and the Congress of South African Trade Unions.

Situating academic discourses within broader discursive domains: the case of legal academic writing

Ruth Breeze
Universidad
de Navarra
rbreeze@unav.es

1 Introduction

Corpus linguistics has frequently been used to research academic language and genres. Since Hyland (2004), a considerable volume of research has been published centring on disciplinary discourses, usually accompanied by explanations as to why differences between disciplines exist, framed in terms of paradigms (Hyland 2004) or values (Giannoni 2011). This paper takes an innovative approach to the issue of disciplinary discourses in academia, by considering the situated nature of academic discourse within a wider discursive domain. In this paper I explore the overlap between academic writing in law and other argumentative legal discourses, in order to establish how much of the specificity of legal academic discourse can be accounted for by the notion of a common legal discourse or register stretching beyond academia into the professional world. By comparing academic research papers in law with research papers from the area of business and management on the one hand, and with legal judgments and opinions on the other, I aim to delineate the overlap between legal academic writing and the discourses of the law, and between legal academic writing and related academic discourses.

2 Material and method

Three 500,000 word corpora were created: Corpus A, containing academic law articles; Corpus B, consisting of academic articles from business and management journals; and Corpus C, made up of judgments and judicial opinions. WordSmith 5.1 and SketchEngine were used to perform word counts, identify lexicogrammatical features and find bundles. The following features were investigated:

- Speech act verbs
- Epistemic verbs, adverbs and adjectives
- Amplifiers and downtoners
- Evaluative adjectives
- Lexical bundles
- Presence of "if".
- Presence of negatives: "not", "never".
- Modals and auxiliary verbs.

3 Results

For reasons of space, only the most salient findings will be outlined in this section. Let us consider first the commonality identified between the two legal corpora, A and C. After this, the similarities between the two academic corpora, A and B, will be briefly outlined.

The first major area of overlap between A and B was that of speech act verbs. For the sake of simplicity, these were classified using the taxonomy devised by Wierzbicka (1987), and it was notable that the two legal corpora coincided in having a high frequency of verbs in the classes of "assert" and "permit", a trend that was not present in corpus B.
Regarding expressions of epistemic certainty, A and C again coincided to a significant degree. However, in the area of epistemic likelihood, no such pattern emerged. As far as modal verbs were concerned, A and C both showed a predilection for the use of "shall", "must" and "could", which were infrequent in the business corpus.

The two legal corpora also both used more downtoners, and fewer amplifiers, than did the business corpus.

As far as formulaic language was concerned, academic law articles (A) and judgments (C) were similar to each other, and contrasted with business articles (B), in the frequency of bundles with an "if" meaning (in the event of, in the case of). Further study of the word "if" itself revealed another striking area of overlap between A and C: "if" was at least twice as frequent in A and C as in B, and conditionals such as those marked by the combination "if...had" were at least four times as frequent in A and C compared to B.

Finally, the two legal corpora each contained at least twice the number of negative constructions (marked by the presence of "not", "no" and "never") when compared to the business corpus.

Commonality between A and B was generally less notable, but we found that they shared a high frequency of speech act verbs belonging to the class of "summing up" and "ordering", and lower frequencies of "forbidding" and "arguing", than corpus C. Corpus C was also found to have a far higher frequency of bundles with a referential function than did A and B, particularly those used to refer to legislation and rules ("in accordance with the", "within the meaning of", "in the light of"), and those used to indicate purpose or result ("for the purpose of", "as a result of").

4 Discussion

On the basis of these findings, it seems fair to say

that an analysis of legal academic writing under the microscope of corpus linguistics places it close in many respects to other discursive legal genres (judgments and opinions) than academic writing from neighbouring fields. Although this would be expected in the area of lexis, particularly discipline-specific terminology, sub-technical terms and so on, it is less obvious why this would be so in the choice of speech act verbs or the use of amplifiers or downtoners. The reason may lie in the way legal professionals construct themselves discursively, which determines the speech acts they use even in non-judicial settings.

The results in the area of grammatical constructions, particularly the predominancy of conditionals and negatives in legal genres, are particularly interesting, since they point towards other significant features of legal discourse that appear to cross the boundaries between academia and the courtroom. Legal discourse is essentially polyphonic, the main line of argument being constructed carefully against a number of alternative views or interpretations. The frequency of negatives points to the need to rule out other possible arguments, while the use of conditionals is probably a product of the use of dialogic argumentation involving the consideration (and refutation) of different voices and different constructions of both argument and fact.

Returning to the original research question, we can see that corpus linguistics provides evidence of the embeddedness of one particular academic discourse within the broader professional discourses of the field. Contrastive research on academic discourse would benefit from broadening its scope to take in the genre systems of the discipline, rather than focusing on academic articles in isolation.

5 Acknowledgements

The research for this paper was carried out within the framework of the project "Metadiscurso y lenguaje evaluativo: perspectivas teóricas y de análisis en el discurso periodístico", funded by the Spanish Ministerio de Economía y Competitividad (ref. FFI2012-36309).

References

Giannoni, D., 2011. *Mapping academic values in the disciplines: a corpuss-based approach*. Bern: Peter Lang.

Hyland, K., 2004. *Disciplinary discourses: social interactions in academic writing*. Ann Arbor: University of Michigan.

Wierzbicka, A. 1987. *English speech act verbs: a semantic dictionary*. Sydney: Academic Press.

Collocations in context: A new perspective on collocation networks

Vaclav Brezina
Lancaster University
v.brezina@lancaster.ac.uk

Tony McEnery
Lancaster University
a.mcenery@lancaster.ac.uk

Stephen Wattam
Lancaster University
s.wattam@lancaster.ac.uk

1 Introduction: Collocation networks

The idea that a text in a particular field of discourse is organised into lexical patterns which can be visualised as networks of words that collocate with each other was proposed by Phillips (1985) and later explored in a number of studies using both general and specialised corpora (e.g. Alonso et. al. 2011; McEnery, 2006; Williams, 1998). This idea has very important theoretical implications for our understanding of the relationship between the lexis and the text, and ultimately between the text and the discourse community/the mind of the speaker. Although the approaches so far have offered different possibilities of constructing collocation networks, they have, in our opinion, not yet successfully operationalised some of the desired features of such networks.

This study revisits the concept of lexical/collocation networks and its different operationalizations in corpus linguistics. It lays theoretical groundwork for identification of collocations in a larger context and shows meaningful applications of the notion of collocation networks in a case study on the moralistic discourse around *swearing* in the 17th and 18th centuries. We also introduce *GraphColl*, a new tool for building collocation networks that implements a number of desirable statistical measures that help identify collocation networks.

2 Method

The case study is based on *The Society for the Reformation of Manners Corpus (SRMC)* compiled by McEnery for his 2006 study on swearing in English. In the case study, we replicate McEnery's research and show how the data can be further explored using *GraphColl* and what new insights about the moralistic discourse we can get with the new technique. Table 1 provides an overview of the *SRMC*.

Text	Tokens	Date
Yates	43,016	1699
Walker	63,515	1711
Anon	4,201	1740
Penn	9,800	1745
TOTAL	**120,532**	

Table 1: *Society for the Reformation of Manners Corpus*

The study uses *GraphColl*, a new tool developed by the authors, which builds collocation networks on the fly and gives the user full control over the process of identification of collocations. Our starting node (i.e. the word which we searched for first) was "swearing". The procedure consisted of the following steps:

1 Replication of McEnery's (2006) study – MI2 association measure.
2 Checking the results with log likelihood, another association measure, which looks at the evidence in the data against the null hypothesis.
3 Adding directionality as another dimension of the collocational relationship using directional association measure Delta P (Gries, 2013).
4 Adding dispersion with Cohen's D (Brezina, in preparation).

3 Results and discussion

The study shows how different association measures (i.e. statistics for identification of collocations) highlight different aspects of the moralistic discourse. The following graph displays the results of the replication of McEnery's (2006) study. The highlighted items were discussed in McEnery (2006).

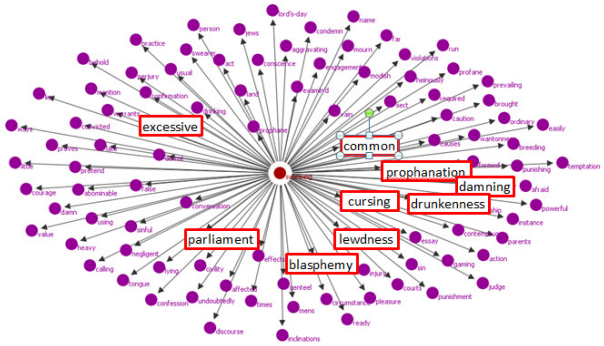


Figure 1: Collocates of "swearing": Replication of McEnery (2006) – 3a-MI2(3), R5-L5, C5-NC1; function words removed

We can see that in addition to the collocates discussed by McEnery, the moralistic discourse on swearing included a number of other associations. In particular, we identified two crucial areas:

- collocates that illuminate the strong religious context of the debate: e.g. *prophane/profane, vain, sinful, conscience, sin* (against god), *damn, condemn* and *Jews.*
- collocates with general negative associations: *dismal, drinking* (as another "sinful" activity), *false, contemptuous, abominable* and *wantonness*

Figure 2 shows the first-order collocation network around the node "swearing" with the log-likelihood as the association measure used. As in Figure 1, the highlighted items mark the overlap with McEnery (2006). Unlike the effect-size measures such as MI2 (see Figure 1), log likelihood tests the amount of evidence in the data against the null hypothesis. In other words, the question we are asking is not how large is the attraction between the node and the collocates, but rather whether we have enough evidence in the data to reject the null hypothesis.

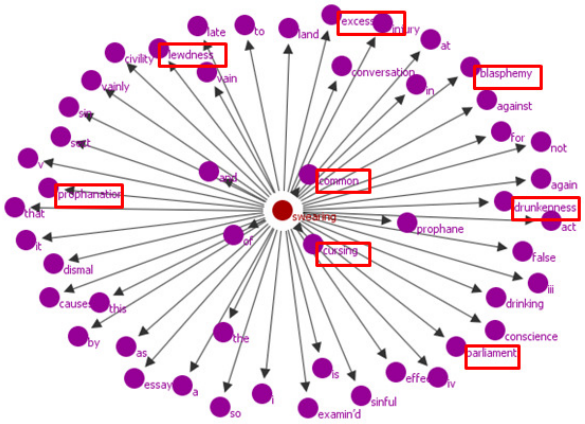


Figure 2: Collocates of "swearing": 6a-LL (15.13), R5-L5, C1-NC1; no filter applied

Although the graph above does not identify a new semantic dimension, it confirms the centrality of those collocates discussed previously, and provides further evidence for the key themes of the moralist debate against swearing – which are 1) connection to other vices (especially drinking and 2) religion

If we want to see further dimensions of the moralistic discourse around swearing, we need to employ a directional association measure Delta P and move beyond the first-order collocates. The analysis is shown in Figure 3.

As we can see, swearing is symmetrically connected with collocates such as *vain, common, cursing* and *prophane*. Interestingly, the noun derived from the adjective *prophane, prophanation*, has a stronger relationship with swearing than vice versa. This means that *prophanation* would more readily trigger the association with *swearing* than *swearing* would with *prophanation*.

Swearing is also connected through *cursing* (its

strongest collocate) to *drunkenness* and (yet again) *prophanation* and through these in turn to a host of other associations including the people who would be referred to as "prophaners". These would be *swearers*, *drunkards* and *(lewd) persons*. In this collocation network we can thus readily see how the abstract moralist discourse evolves and becomes personalised, with its metaphorical finger pointing to specific offenders.

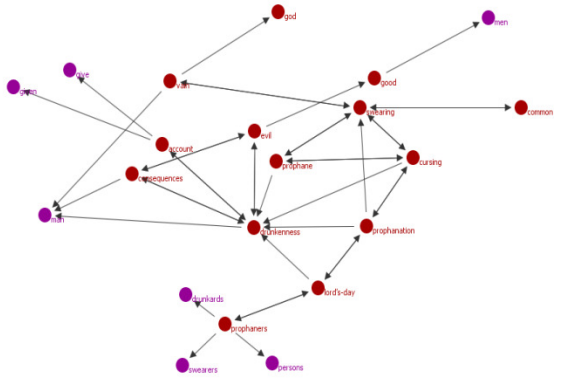


Figure 3: Collocates of "swearing": 13a-Delta P (0.1), R5-L5, C1-NC4; function words removed

Finally, a new association measure, Cohen's d, is briefly discussed. Cohen's d (Algina et al., 2005; Cohen, 1988) is a commonly used measure of effect size outside of corpus linguistics. Here we demonstrate how Cohen's d can be implemented as an association measure which takes into account the distribution of collocates in different texts (or subcorpora) by comparing the mean values of collocate frequencies in the collocation window and outside of the window (see Brezina, in preparation).

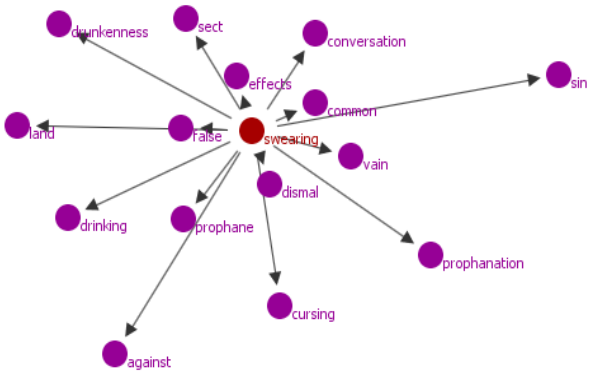


Figure 4: Collocates of "swearing": 14-Cohen's d (0.5), R5-L5, C1-NC4; no filter

Figure 4 shows the collocates of *swearing* in the SRMC identified using Cohen's d. Even with a very new metric, we obtained a stable set of collocates including *cursing*, *drunkenness*, *common* and *vain*. This is a very important signal that the collocational relationship – and collocation networks in particular – are based on the reality of discourse as reflected in language corpora, rather being a function of any particular statistical procedure.

4 Conclusion

The purpose of this study was to demonstrate that connectivity between collocates is an important dimension of the collocational relationship. In a case study, we showed how the collocation networks can be built around the nodes that we are interested in and how these can, in turn, shed more light on word associations in texts and discourse.

References

Algina, J., Keselman, H., & Penfield, R. D. (2005). "An Alternative to Cohen's Standardized Mean Difference Effect Size: A Robust Parameter and Confidence Interval in the Two Independent Groups Case". *Psychological methods, 10*(3), 317.

Alonso, A., Millon, C., & Williams, G. 2011. "Collocational Networks and their Application to an E-Advanced Learner's Dictionary of Verbs in Science (DicSci)". Proceedings of eLex, 12-22.

Brezina, V., McEnery, T. & Wattam, S. (under consideration), "Collocations in context: A new perspective on collocation networks", *International Journal of Corpus Linguistics.*

Cohen, J. (1988). *Statistical power analysis for the behavioral sciencies*. Hillsdale: Lawrence Erlbaum Associates.

Gries, S. T. 2013. "50-something years of work on collocations: what is or should be next" *International Journal of Corpus Linguistics*, 18(1), 137-166.

McEnery, T. 2006. *Swearing in English: Bad language, purity and power from 1586 to the present*. Abington, Oxon: Routledge.

Phillips, M. 1985. *Aspects of text structure: An investigation of the lexical organisation of text*. Amsterdam: North-Holland.

Williams, G. 1998. "Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles." *International Journal of Corpus Linguistics*, 3 (1), 151-171.

A Corpus analysis of discursive constructions of the Sunflower Student Movement in the English Language Taiwanese press

Andrew Brindle
St. John's
University
andrewbr@mail.sju.edu.tw

1 Introduction

On March 18, 2014, student protesters in Taipei stormed the Legislative Yuan, Taiwan's chamber of parliament, beginning a 24-day occupation which paralysed the island's legislature (Fukuda 2014). The protests, driven by a coalition of students and civic groups, later given the name Sunflower Student Movement, were in response to the ruling Kuomintang's (KMT) attempt to unilaterally ratify a trade pact, the Cross-Strait Service Trade Agreement (CSSTA), with the People's Republic of China (PRC). Under the terms of the treaty, service industries such as banking, healthcare, tourism, telecommunication and publishing would be opened to investment. Protesters perceived the pact would be detrimental to the Taiwan economy and leave it vulnerable to political pressure from Beijing.

This study examines the discursive constructions of the Sunflower Student Movement in the two English language newspapers in Taiwan. News discourse and media language is of relevance to social scientists due to its omnipresence in contemporary society, the public attention it receives and the political influence it generates (Mautner 2008). Fairclough (1995) and van Dijk (1988) consider the media as key ideological brokers, able to reproduce and maintain discourses of dominant social order. Thus, news discourse may be considered as ideological and a significant influence within society (Costelloe 2014). The journalistic response to certain issues is of vital consequence to how the public comprehend and participate in socio-political events such as the Sunflower Student Movement. Therefore, the study of newspaper discourses of anti-government protests in Taiwan, may facilitate the comprehension of ideological discourses prevalent within Taiwanese society.

2 Data

The data were collected from the online editions of the two English language daily newspapers in Taiwan, *The China Post* and the *Taipei Times*. *The China Post* was established in 1952 and claims to have a daily readership of over 400,000 through online and print media. The newspaper was established during the martial law era of Taiwan (1949-1987), a time in which the media was controlled by the KMT, resulting in a culture of deference toward the government and emphasis of Han Chinese identity over that of Taiwanese (Rawnsley 2004). Thus, the newspaper is seen as supporting a nationalist, pro-unification agenda. The *Taipei Times* was established in 1999 and has a pro-Taiwan independence editorial stance (Kuo 2007). The newspaper claims to have the largest circulation of Taiwan's English language newspapers and according to its website, the online edition receives approximately 200,000 hits a day.

In a 6-month period beginning the day the legislature occupation began (March 19 – August 30, 2014), all articles which contained the search terms `student protest` or `Sunflower Movement` were collected. The articles were checked before being added to the corpora in order to verify that they were related to the protests taking place in Taipei and not another protest elsewhere. The China Post corpus consisted of 245 articles, 122,633 words; the Taipei Times corpus comprised of 285 articles, 187,717 words.

3 Findings

In this study, a corpus-based discourse analysis (Baker 2006) was undertaken examining lexical frequency and keywords within the corpora, thereby considering the emerging patterns of the discursive construction of the student protests in the two newspapers. A focus on frequency and keywords combined with concordance and collocation analysis can provide helpful indications of the ideological stance of the newspapers, which may in turn reflect the opinions held by the readership (van Dijk 1991).
A preliminary analysis of lexical, non-function word frequency shows the most frequent words as following:

The China Post – `Taiwan` (freq. 806, 6572 per million), `Yuan` (freq. 562, 4582 per million), `students` (freq. 552, 4501 per million), `protesters` (freq. 437, 3563 per million), `Legislative` (freq. 419, 3416 per million).
Taipei Times – `Taiwan` (freq. 1149, 6120 per million), `movement` (freq. 737, 3926 per million), `students` (freq. 708, 3771 per million), `Ma` (freq. 662, 3526 per million), `trade` (freq. 655, 3489).
The word `Taiwan` is the most frequent in both corpora, however, collocates demonstrate different stances. In *The China Post*, collocates of `Taiwan` include: `economy`, `development`, `competitiveness`, `industry`, `hurt` and `affect`. In the *Taipei Times*, collocates include:

safeguard, protect, defend, support, democracy, future, sovereignty and independence. Such findings appear to indicate that *The China Post* constructs the protests in terms of the damage they may cause to the status quo and economic stability of the island, whereas the discursive strategy of the *Taipei Times* presents the protests as a struggle to protect and defend the sovereignty and independence of the island as well as safeguarding the democratic process. The word students is also one of the most frequent words in both corpora; in *The China Post*, collocates include storm, evict, urge and demand, however, in the *Taipei Times* support, occupy and participate are collocates, thus indicating that while one newspaper focuses on the violent nature of the demonstrations, the other emphasises solidarity with the students. Such discursive constructions are further perpetuated when extended frequency lists are analysed. A frequent word in the China Post corpus is police with a focus on the violence which occurred between the protesters and police during the protests, thus emphasising the anti-social nature of the demonstrations. Such a discourse of violence is absent from the Taipei Times data; a high frequency word is Sunflower which not only functions as a nominalisation strategy associated with hope, but also associates the protest movement with the Wild Lily movement, a student movement in 1990, which marked a turning point in Taiwan's transition to pluralistic democracy.

Following the study of frequency, keywords of the corpora were analysed, firstly using enTenTen (2012) as a reference corpus.

	The China Post Corpus			*The Taipei Times Corpus*		
#	*keyword*	*score*	*freq.*	*keyword*	*score*	*freq.*
1	DPP	1889	321	KMT	1753	398
2	KMT	1517	225	DPP	1107	288
3	pact	996	461	Sunflower	781	438
4	Kuomintang	719	103	Taiwanese	774	352
5	Taiwan	685	806	Taiwan	638	1150
6	Taipei	533	158	pact	631	447
7	protesters	491	454	Taipei	613	278
8	Tsai	487	78	Tsai	318	78
9	Sunflower	467	171	protesters	299	423
10	Jiang	430	168	Jiang	246	142

Table 1: Keywords ordered by keyness

The keywords with the highest keyness scores are similar in both corpora with the exception of Kuomintang in the China Post corpus and Taiwanese in the Taipei Times corpus. However, by studying collocates of the keywords, differing discursive constructions emerge. An example of this is the analysis of collocates of KMT, the pro-unification governing party. In *The China Post*, KMT collocates with words such as: enjoy, agree, attempt, reiterate, propose, support, join and present. In the *Taipei Times*, KMT collocates with: underestimate, betray, accuse, refuse, embezzle, strip, kill, damage and suffer. Thus it can be seen how the discursive constructions differ.

When the two corpora are compared against each other, the keywords with the highest level of keyness in the China Post corpus when the Taipei Times corpus is used as a reference are: services, Kuomintang, hall, activists, Mainland, assembly, police, ruling, parliament and Yuan. When the Taipei Times corpus is studied with the China Post corpus as a reference, the keywords with the highest level of keyness are: Chinese, Taiwanese, Sunflower, political, democracy, system, movement, constitutional, Government's, and handling. When concordance lines of these words are studied, discursive patterns emerge, as the following examples indicate:

The China Post

Should the **Sunflower student movement** succeed in getting the trade agreement retracted, it would be an economic disaster.
Student activists proceeded to tear down the doors, destroy voting devices installed in the building and empty the drawers of several lawmakers' desks
In fact, there has been a clear shift from overwhelming attention on the **student protesters** and toward more growing coverage of counter voices.

Taipei Times

From all walks of life, supporters of the **"Sunflower student movement"** took to the street in Taipei yesterday, marked by festivity, diversity and order.
Taiwanese are fully aware of the course of events, and many in academia, civil society and among

the public have expressed support for the **student protesters.**
It would come as no surprise if some of the **student leaders** of today become the legislative leaders of tomorrow.

4 Conclusion

The findings demonstrate that the *Taipei Times*, associated the movement with democracy movements from the past, while constructing the protests as a struggle to uphold democracy and Taiwanese independence, and furthermore emphasised the support the movement received from the general public. *The China Post* constructed the protests negatively, focusing on the destabilising elements of the protests such as the economic consequences of the occupation, instances of violence, disruption to the status quo, as well as constructing the protesters as being unrepresentative of the general population of the island. The differing discursive constructions of the protest movement may reflect divisions within the Taiwanese society in relation to questions of nationhood, independence and its stance towards the increasing economic and political influence of the PRC.

References

Baker, P. 2006. *Using Corpora in Discourse Analysis.* London: Continuum.

Costelloe, L. 2014. Discourses of sameness: Expressions of nationalism in newspaper discourse on French urban violence in 2005. In *Discourse & Society* 2014, Vol. 25(3) 315-340.

Fairclough, N. 1995. *Media Discourse*. London: Edward Arnold.

Fukuda, M. 2014. *Japan-China-Taiwan Relations after Taiwan's Sunflower Movement*. Asia Pacific Bulletin. Number 264.

Kuo, S. 2007. Language as Ideology. Analyzing Quotations in Taiwanese News Discourse. In *Journal of Asian Pacific Communication*. 17:2. 281-301.

Mautner, G. 2008. Analyzing Newspaper, Magazines and other Print Media. In R. Wodak & M. Krzyzanowski (eds.) *Qualitative Discourse Analysis in the Social Sciences*. New York: Palgrave Macmillan.

Rawnsley, G. D. 2004. Treading a Fine Line: Democratisation and the Media in Taiwan. In *Parliamentary Affairs*. Vol. 57, Issue 1. 209-222.

van Dijk, T. 1988. *News as Discourse*. Hillsdale, NJ: Lawrence Erlbaum.

van Dijk, T. 1991. *Racism and the Press.* London: Routledge.

An examination of learner success in UCLanESB's B1 and C1 speaking exams in accordance with the Common European Framework of Reference for Languages.

Shelley Byrne
University of Central Lancashire
sbyrne@uclan.ac.uk

Recent years have seen an increase in the application of learner corpora, for example, the International Corpus of Learner English (Granger et al. 2009), the Cambridge Learner Corpus (Cambridge University Press 2015), English Profile (Cambridge English Profile Corpus n.d.) and the Vienna-Oxford International Corpus of English (VOICE 2013), to obtain greater insight into proficiency and linguistic development during second language learning. In conjunction with definitions of competence (Chomsky 1965; Hymes 1972; Canale and Swain, 1980; Celce-Murcia et al. 1995), continuing attempts are being made to pinpoint which skills and knowledge need to be developed if learners are to 'succeed' in operating in a target language.

One document, the Common European Framework of Reference for Language (CEFR) (Council of Europe [CoE] 2001), details extensively the wide-ranging contexts for language use and the potential abilities to be evidenced by learners at different levels. However, despite its non-prescriptive intentions and its purpose of providing an adaptable *guide* to varying language provision contexts (CoE 2001), it has faced criticism. Whilst some warn of the misapplication of the CEFR and occasional notions assuming a 'gold standard' of language teaching, others highlight its lack of supporting second language acquisition theory and, more significantly for this study, its absence of authentic language use to explicate its six proficiency levels and their illustrative 'can do' statements (see Davidson and Lynch 2002; Fulcher 2004; Weir 2005; Alderson et al. 2006; Alderson 2007; Hulstijn 2007; Little 2007; Jones and Saville 2009; Fulcher et al. 2011).

With the CEFR and its levels being valuable tools in the field of language assessment and test design (Coste 2007; Little 2007; Jones and Saville 2009), this study aims to examine what makes spoken language use at B1 and C1 *successful* in the University of Central Lancashire's English Speaking Board Exams. Using a learner corpus of B1 spoken test data (22740 words) and a learner corpus of C1 spoken test data (26620 words) based solely on candidates receiving a clear pass, the study aims to

provide clarification of what lexico-grammatical competence may comprise at these levels, which CEFR descriptors occur in the B1 and C1 speaking tests and how 'can do' statements may be realised. The findings stemming from corpus tools including vocabulary profiles, word frequency lists, keyword lists and three- and four-word lexical chunk analysis, as well as a qualitative investigation of 'can do' occurrence, also endeavour to highlight shared or diverging features of successful language use in the specified B1 and C1 speaking tests.

References

Alderson, J. C. 2007. "The CEFR and the need for more research". *The Modern Language Journal*. 91 (4): 659-663.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S. and Tardieu, C. 2006. "Analysing tests of reading and listening in relation to the common European framework of reference: The experience of the Dutch CEFR construct project". *Language Assessment Quarterly: An International Journal*. 3 (1): 3-30.

Cambridge English Profile Corpus. (n.d.). English Profile: CEFR for English. Available online at http://www.englishprofile.org/index.php/corpus

Cambridge University Press. 2014. Cambridge English Corpus. Available online at http://www.cambridge.org/about-us/what-we-do/cambridge-english-corpus

Canale, M. and Swain, M. 1980. "Theoretical bases of communicative approaches to second language teaching and testing". Applied Linguistics, 1: 1-47.

Celce-Murcia, M., Dörnyei, Z. and Thurrell, S. 1995. "Communicative competence: A pedagogically motivated model with content specifications". *Issues in Applied linguistics*, 6 (2): 5-35.

Chomsky, N. 1965. Aspects of the theory of syntax. Cambridge, MA: MIT Press.

Coste, D. 2007. "Contextualising uses of the common European framework of reference for languages". In *Report of the intergovernmental Forum, the Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities*.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, Cambridge University Press.

Davidson, F. and Lynch, B. K. 2002. *Testcraft: A teacher's guide to writing and using language test specifications*. Yale University Press.

Fulcher, G. 2004. "Deluded by artifices? The common European framework and harmonization". *Language Assessment Quarterly: An International Journal*. 1 4: 253-266.

Granger, S., Dagneaux, E., Meunier, F. and Paquot, M. 2009. ICLE. Available online at http://www.uclouvain.be/en-cecl-icle.html

Hulstijn, J. H. 2007. "The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language Proficiency". *The Modern Language Journal*. 91 (4): 663-667.

Hymes, D. 1972. *On communicative competence. In J. B. Pride and J. Holmes* (eds.) Sociolinguistics. Middlesex: Penguin.

Jones, N., and Saville, N. 2009. "European language policy: Assessment, learning, and the CEFR". Annual Review of Applied Linguistics. 29: 51-63.

Little, D. 2007. "The Common European Framework of Reference for Languages: Perspectives on the Making of Supranational Language Education Policy". *The Modern Language Journal.* 91: 645.

VOICE. 2013. VOICE: *Vienna-Oxford International Corpus of English*. Available online at https://www.univie.ac.at/voice/page/what_is_voice

Weir, C. J. 2005. Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*. 22 (3): 281-300.

Automated processing, grading and correction of spontaneous spoken learner data

Andrew Caines
University of Cambridge
apc38@cam.ac.uk

Calbert Graham
University of Cambridge
crg29@cam.ac.uk

Paula Buttery
University of Cambridge
pjb48@ cam.ac.uk

Michael McCarthy
University of Cambridge
mactoft@cantab.net

1 Overview

A research area of growing interest lies at the intersection of computational linguistics and learner corpus research: automated language teaching and assessment. We describe the work we have carried out to date in our group, including syntactic and phonetic analyses, sentence boundary detection, and error annotation.

The key contribution of our work is in the methods, tools and frameworks we are developing to work with *spontaneous* learner speech — for example, a new annotation framework for error analysis that treats errors and infelicities in a gradient and multi-dimensional way (Buttery et al 2014). Much of these developments are built on existing research and technology; for example, the aforementioned error analysis framework operationalizes the *accuracy* component of the Complexity-Accuracy-Fluency framework (Housen & Kuiken 2009).

For now, we work with learner test data from Cambridge English Language Assessment's *BULATS* business English oral exam, which contains both scripted and unscripted sections. We focus on the latter, and investigate the effects of proficiency, age and first language.

2 Automated assessment

In order to support and streamline existing learner assessment by human examiners, various automatic grading systems have been developed by several different groups. For example, the Educational Testing Service (ETS) provide the *Criterion*® *Online Writing Evaluation* service to learners via teaching institutions, an application based on their *e-rater*® scoring engine (Burstein 2003).

Microsoft Research (MSR) offered a web service called *ESL Assistant* (Gamon et al 2009) which has now been withdrawn from service, although MSR state that the error detection and correction components remain to be used as required.

Meanwhile, researchers at the University of Cambridge developed the *SAT* (self-assessment and tutoring) *system* for intermediate learners of English (Andersen et al. 2013). The *SAT System* is now freely available on the iLexIR website[27], presented as *Cambridge English Write & Improve,* and offers automated assessment and corrective feedback on user essays.

These systems have so far been set up to work with written texts. We aim to develop a comparable system for spoken language.

3 Automatically grading speech

Previous efforts at assigning proficiency grades to learner speech have typically involved language models and automatic speech recognition (ASR) within scripted or constrained lexical domains (de Wet et al 2009, Cucchiarini et al 2014).

We intend to build on this work in the less restricted domain of unplanned spoken data (albeit responding to examiner prompts), while incorporating further linguistic features, making use of assessment algorithms from the *SAT System*, and allowing for our other ultimate goal — the automatic provision of linguistically-meaningful learner feedback — at all stages of system design.

However, first of all we must tackle the not inconsiderable problem of the automated processing of free speech.

4 NLP for spoken language

Natural language processing (NLP) technology has for the most part been developed and trained on the basis of standard English texts written by native speakers, and as a result is well equipped to deal with unseen data of this type (Manning & Schütze 1999). There are ongoing parallel efforts to adapt this technology to other domains such as internet and non-native corpora (Kilgarriff & Grefenstette 2003; Leacock et al 2014), and other languages (Abney & Bird 2010).

Meanwhile, it's fair to say that NLP for spoken language is less advanced, with the several challenges inherent to this medium now presented to us. Firstly, the speech must be transcribed for use by NLP tools. In our project this is addressed via crowdsourcing and ASR.

Secondly, once the transcriptions are in hand, it must be segmented into usable chunks — 'sentences', as it were (to make use of this writing-specific term in its loosest sense). This task of sentence boundary identification is a work-in-progress for us, but we will show how automated methods compare to a manually-assigned 'gold

[27] https://sat.ilexir.co.uk

standard' of sentence boundaries, as well as sentence boundaries naively based on the length of silent pauses.

In addition, features typical of unplanned speech, such as filled pauses, false starts, ellipsis, and verbless utterances, must be dealt with, either by 'cleaning' them up so that the text is more written-like, or by incorporating them into the syntactic analysis via specialised tree structures. We show the pros and cons of both of these approaches, and weigh up the philosophical argument that speech should be left as is and treated as a fully valid medium in its own right, versus the practical consideration that adaptation to written language allows immediate use of existing NLP tools.

5 Computer-aided language learning

Once we have the data in a usable format, we next need to classify the learner's proficiency level, in the process identifying errors and extracting features of linguistic interest with which we can offer corrective feedback.

This part of the system relates to computer-aided language learning (CALL) and is the end-goal of our work. We show what kind of linguistic features we are working with — both phonetic and morph-syntactic — for example, vowel realisations and agreement errors — and we discuss the kind of feedback we can give.

We plan to empirically test the effectiveness of such feedback, and we fully intend that our CALL system will be interactive and individualised.

Acknowledgements

This work has been funded by Cambridge English Language Assessment. We thank Nick Saville and Ted Briscoe for their guidance. We thank Francis Nolan, Kate Knill, Rogier van Dalen, and Ekaterina Kochmar for their help. And we gratefully acknowledge the support of Alan Little, Barbara Lawn-Jones and Luca Savino.

References

Abney, S. & S. Bird (2010). The Human Language Project: Building a Universal Corpus of the World's Languages. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics.

Andersen, Ø., H. Yannakoudakis, F. Barker, & T. Parish (2013). Developing and testing a self-assessment and tutoring system. *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications.* Association for Computational Linguistics.

Burstein, J. 2003. The *e-rater*® scoring engine: automated essay scoring with natural language processing. In: M.D. Sheers & J. Burstein (eds.) *Automated essay-scoring: a cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Buttery, P.J., A.P. Caines & M.J. McCarthy (2014). Infinite shades of grey: what constitutes an error? Presentation at the *IVACS Conference 2014,* Newcastle.

Cucchiarini, C., S. Bodnar, B. Penning de Vries, R. van Hout & H. Strik (2014). ASR-based CALL systems and learner speech data: new resources and opportunities for research and development in second language learning. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* European Language Resources Association.

de Wet, F., C. Van der Walt & T.R. Niesler (2009). Automatic assessment of oral language proficiency and listening comprehension. *Speech Communication* 51: 864-874.

Gamon, M., C. Leacock, C. Brockett, W.B. Dolan, J. Gao, D. Belenko & A. Klementiev (2009). Using statistical techniques and web search to correct ESL errors. *CALICO Journal* 26: 491-511.

Housen, A. & F. Kuiken (2009). Complexity, fluency and accuracy in second language acquisition. *Applied Linguistics* 30: 461-473.

Kilgarriff, A. & G. Grefenstette (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics 29*(3).

Leacock, C., M. Chodorow, M. Gamon & J. Tetreault (2014). *Automated Grammatical Error Detection for Language Learners,* 2nd edn. San Rafael, CA: Morgan & Claypool.

Manning, C. & H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

A longitudinal investigation of lexical bundles in a learner corpus

Duygu Candarli
University of Manchester
duygu.candarli
@postgrad.manchester.ac.uk

1 Introduction

Phraseology has been a major area of interest within the fields of English for Specific Purposes and English for Academic Purposes. There is a growing body of work on phraseological patterns in published research articles from both cross-cultural and cross-linguistics perspectives. One type of phraseological patterns are lexical bundles that have been well-documented in terms of frequency, use, structure, and discourse functions in expert academic writing (Biber et al. 2004; Biber 2009; Cortes 2004, 2013).

In novice academic writing, previous studies mainly followed the cross-sectional research design in that data were collected at a single point in time. More recent studies have investigated the use of lexical bundles in foreign/second language writing across different levels (Ädel and Römer 2012; Chen and Baker 2014; Staples et al. 2013). Though these pseudo-longitudinal studies provided valuable insights into the characteristics of phraseology across different proficiency levels, it would be worthwhile to capture the developmental and interlanguage features of learner writing within a truly longitudinal research design. In recent years, longitudinal research on L2 phraseological patterns of novice academic writers in an immersion setting has been increasing (Bestgen and Granger 2014; Li and Schmitt 2009). Nevertheless, little is known about the phraseological development of language learners/users in an EFL setting. The present study addresses these two following questions:

- To what extent, if any, does the frequency of lexical bundles change in the essays of Turkish learners of English over one academic year?
- To what extent, if any, do the discourse functions of lexical bundles change over one academic year?

2 Data

The learner corpus consists of 300 English essays of 100 Turkish students who were in their first year at an English-medium university in Turkey. Each essay is approximately 500 words in length. The essays were collected at the beginning of the first semester, at the end of the first semester, and at the end of the second semester from the same students.

The participants had received high scores in the English language section of the university entrance examination before entering the university. Furthermore, before embarking on their studies, they had to pass the English language proficiency test of the university with a good score which is the equivalent of an overall band of 6.5 in IELTS (Academic) with no less than 6.5 in writing skill. Students could also submit their IELTS (at least 6.5) or TOEFL IBT (at least 79) test reports. In their first year at the university, they take English language courses to brush up on their English language skills. They also take 'Advanced Writing in English' courses at both fall and spring semesters in their first year. The students submit all their assignments in English during their undergraduate education except two Turkish language courses. It could be said that the students were expected to internalise academic discourse and begin academic socialisation through academic writing. As Ortega and Iberri-Shea (2005) stated, longitudinal research in language learning is "better motivated when key events and turning points in the social or institutional context investigated are considered" (p. 38). Though one year may not be long enough to regard this study as longitudinal, it was designed to offer insights into Turkish EFL students' phraseological development.

3 Methodology

The current study employed a corpus-driven approach in that the analysis was based on the most frequent multi-word units. Previous studies mostly focused on four-word lexical bundles. In this study, a more inclusive approach was taken, and three-, four- and five-word sequences were examined in terms of frequency, use and discourse functions. The free concordance program AntConc (version 3.4.1) was used to extract lexical bundles (Anthony 2014). As the corpus was small, the frequency threshold was set to 10 times per million. Regarding the dispersion criterion, a sequence had to occur in at least five different texts in the corpus (see Biber et al. 2004; Conrad 2013). Following Chen and Baker (2014), I refined the lexical bundles when there were overlapping word sequences and partial subsumption. Moreover, Biber et al.'s taxonomy (2004) was adapted to categorise the discourse functions of lexical bundles.

4 Preliminary findings

The preliminary findings revealed that there was very little, if any, change in the functional distribution of lexical bundles in the learner corpus over one year. However, the frequency of lexical bundles slightly decreased over one academic year. The present findings seem to be consistent with

other research which found little change in the longitudinal development of phraseology (Li and Schmitt 2009) and little difference across different proficiency levels (Staples et al. 2013). These results suggest that the learners might rely on lexical bundles to a lesser extent as they gain experience in academic writing. Further research could investigate lexical frames which can give a detailed picture of pattern variability in learner corpora.

References

Ädel, A. and Römer, U. 2012. "Research on advanced student writing across disciplines and levels: Introducing the Michigan corpus of upper-level student papers". *International Journal of Corpus Linguistics* 17 (1): 3–34.

Anthony, L. 2014. *AntConc (Version 3.4.3)* [Computer Software] Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Bestgen, Y. and Granger, S. 2014. "Quantifying the development of phraseological competence in L2 English writing: An automated approach". *Journal of Second Language Writing* 26: 28–41.

Biber, D. 2009. "A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing". *International Journal of Corpus Linguistics* 14 (3): 275–311.

Biber, D., Conrad, S. and Cortes, V. 2004. "If you look at . . . : Lexical bundles in university teaching and textbooks". *Applied Linguistics* 25 (3): 371–405.

Chen, Y.-H. and Baker, P. 2014. "Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1". *Applied Linguistics*, 1–33.

Cortes, V. 2004. "Lexical bundles in published and student disciplinary writing: Examples from history and biology". *English for Specific Purposes* 23 (4): 397–423.

Cortes, V. 2013. "The purpose of this study is to: Connecting lexical bundles and moves in research article introductions". *Journal of English for Academic Purposes* 12 (1): 33–43.

Li, J. and Schmitt, N. 2009. "The acquisition of lexical phrases in academic writing: A longitudinal case study". *Journal of Second Language Writing 18* (2): 85–102.

Ortega, L. and Iberri-Shea, G. 2005. "Longitudinal research in second language acquisition: Recent trends and future directions". *Annual Review of Applied Linguistics* 25: 26-45.

Staples, S., Egbert, J., Biber, D. and McClair, A. 2013. "Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section". *Journal of English for Academic Purposes* 12 (3): 214-225.

Linguistic preprocessing for distributional analysis efficiency: Evidence from French

Emmanuel Cartier
Université Paris 13
Sorbonne Paris Cité
emmanuel.cartier@lipn.univ-paris13.fr

Valeriya Vinogradova
Université Paris 13
Sorbonne Paris Cité
valeriya.vinogradova@gmail.com

1 Introduction: Distributional Hypothesis and Cognitive Foundations

For about fifteen years, the statistical paradigm, from the distributionalism hypothesis (Harris, 1954) and corpus linguistics (Firth, 1957), has prevailed in the NLP field, with a lot of convincing results: multiword expression, part-of-speech, semantic relation identification, and even probabilistic models of language. These studies have identified interesting linguistic phenomena, such as collocations, "collostructions" (Stefanowitsch, 2003), « word sketches » (Kilgariff et al., 2004).
Cognitive Semantics (Langacker, 1987, 1991; Geeraerts et al. 1994 ; Schmid, 2007, 2013), have also introduced novel concepts, most notably that of « entrenchment », which enables to ground the social lexicalization of linguistic signs and to correlate it with repetition in corpus.

Finally, Construction Grammars (Fillmore et al., 1988 ; Goldberg, 1995, 2003 ; Croft, 2001, 2004, 2007) have proposed linguistic models which reject the distinction lexicon (list of "words") - grammar (expliciting the combination of words) : all linguistic signs are constructions, from morphemes to syntaxical schemes, leading to the notion of "constructicon", as a goal for linguistic description.

2 Computational Models of the Distributional Hypothesis

As Computational Linguistics is concerned, the Vector Space Model (VSM) has prevailed to implement the distributional hypothesis, giving rise to continuous sophistication and several state-of-the-arts (Turney and Pantel, 2010; Lenci and al., 2010; Kiela and Clark, 2013; Clark, 2015). (Kiela and Clarke, 2014) state that the following parameters are implied in any VSM implementation: *vector size, window size, window-based or dependency-based context, feature granularity, similarity metric, weighting scheme, stopwords and high frequency cut-off*. Three of them are directly linked to linguistic preprocessing : *window-based or dependency-based context*, the second requiring a

dependency analysis of the corpus; *feature granularity*, ie, the fact of taking into account either the raw corpus, or a lemmatized or pos-tagged one for n-gram calculus; *stopwords and high frequency cut-off* , ie removal of high-frequency words or "tool words". (Kiela and Clarke, 2014) conducted six experiments/tasks with varying values for each parameter, so as to assess the most efficient ones. They conclude that : dependency-based does not trigger any improvement over raw-text n-gram calculus; as for feature granularity, that stemming yields the better results; as for stopwords or high-frequency words removal, it does yield better results, but *only if* no raw frequency weighting is applied to the results; this is in line with the conclusion of (Bulinaria and Levy, 2012).

Nevertheless, these conclusions should be refined and completed:

1/ As feature granularity is concerned, the authors do not take into account a combination of features from different levels; (Béchet et al., 2012), for example, have shown that combining features from three levels (form, lemma, pos-tag) can result in better pattern recognition for specific linguistic tasks; such a combination is also in line with the Cognitive Semantics and the Construction Grammar hypothesis, that linguistic signs emerge as constructions combining schemes, lemmas and specific forms;

2/ The experiments on dependency need additional experiments, as several works (for example Pado and Lapata, 2007) made a contradictory conclusion.

3/ Stopwords or high-frequency words removal results in better results if no frequency weighting is applied; but the authors apply – as quasi all work in the field -, a brute-force removal either based on "gold standard" stopword lists, or on a arbitrary count to cut off results; this technique should be refined to remove only the noisy words or n-grams and should be linguistically motivated.

3 Linguistic motivation for linguistic preprocessing

The hypothesis supported in this paper is that, if repetition of sequences is the best way to access usage and to induce linguistic properties, language users do not only rely on the sequentiality of language, but also on non-sequential knowledge thus untractable from the actual distribution of words. This knowledge is linked to the three classical linguistical units: lexical units, phrases and predicate structures, each being a combination of the preceding with language-specific rules for their construction. Probabilistic models of language have mainly focused until now on the lexical units level, but to leverage language, probabilistic research must also model and preprocess phrases and predicate structures.

The present paper will try to ground this hypothesis through an experiment, aimed at retrieving lexico-semantic relations in French, where we preprocess the corpus in three ways :

1. morphosyntactic analysis
2. peripheral lexical units removal
3. phrases identification.

As we will see, these steps enable to access more easily the predicate structures that the experiment aims at revealing, while using a VSM model on the resulting preprocessed corpus.

4 Evidence from French: Semantic Relations and Definitions

Definition model: Here we assume that a definitory statement is a statement asserting the essential properties of a lexical unit. It is composed of the definiendum (DFM), i.e. the lexical unit to be defined; the definiens (DFS), i.e. the phrasal expression denoting the essential properties of the DFM lexical unit; the definition relator (DEFREL), i.e. the linguistic items denoting the semantic relation between the two previous elements.

The traditional model of definition decomposes the DFS into two main parts : HYPERNYM + PROPERTIES.

Definitory statement can also comprise other information : enunciation components (*according to Sisley, a G-component is a ...*); domain restrictions (*in Astrophysics, a XXX is a YYY).*

5 Corpus

We use three corpora and retain only the nominal entries in each:

Trésor de la Langue Française (TLF): 61 234 nominal lexical units, and 90 348 definitions;

French Wiktionary (FRWIK): 140 784 nouns, for a total of 187 041 definitions.

Wikipedia (WIKP): 610 013 glosses (ie first sentence of each article) from the French Wikipedia, using a methodology next to (Navigli and al, 2008)

The first two are dictionaries (TLF, FRWIK), the last one is an encyclopedia (WIKP). In the first case, definition obeys to lexicographic standards, whereas definitions are more "natural" in WIKP.

6 System Architecture

The system is composed of four steps:

1 Morpho-syntactic analysis of the corpus

2 Semantic Relation Trigger words Markup

3 Sentence Simplification : this step aims at reducing, as much as possible, the sentences

to the core semantic expressions of definition;

4. Lexico-syntactic pattern-matching for semantic relations : relation(X,Y)

In the following, we will focus on the simplification step.

7 Sentence simplification

Sentence simplification has two main goals :

1. decompose any sentence into its main predicate-arguments structure, and remove and record peripherical elements if necessary;
2. Unify nominal phrases, as they are the target for hypernym relations and their sparsity complicate retrieval of patterns.

Take the following source definition:

en/P cuisine/NC ,/PONCT un/DET DEFINIENDUM être/V un/DET pièce/NC de/P pâte/NC aplatir/VPP ,/PONCT généralement/ADV au/P+D rouleau/NC à/P pâtisserie/NC ./PONCT ((cooking) an undercrust is a piece of dough that has been flattened, usually with a rolling pin.)

It will be reduced to:

un/DET DEFINIENDUM être/V un/DET pièce/NC de/P pâte/NC aplatir/VPP ,/PONCT au/P+D rouleau/NC à/P pâtisserie/NC ou/CC un/DET laminoir/NC ./PONCT

And we extract the domain restriction: *en/P cuisine/NC.*

8 Steps 1 and 2: Adverbials and subordinate clauses

The first linguistical sequences removed from the source sentence are adverbials and specific clauses. But we would like to remove only clauses dependent on the main predicate, not those dependent on one of its core components. For example, we remove the incidential clause in :

DEFINIENDUM (/PONCT parfois/ADV Apaiang/NPP ,/PONCT même/ADJ prononciation/NC)/PONCT être/V un/DET atoll/NC de/P le/DET république/NC du/P+D Kiribati/NPP ./PONCT

But relative clauses dependent on one of the definiens component should be first extracted:

DEFINIENDUM être/V du/P+ enzymes/NC qui/PROREL contrôler/V le/DET structure/NC topologique/ADJ de/P l'ADN/NC ...

To achieve this goal, we use distributional analysis on these clauses with (SDMC, Béchet et al., 2012) and human tuning to determine the most frequent patterns and trigger words of incidential clauses at specific locations in the sentence: beginning of the definition, between the definiendum and a definition relator.

Some incidential clauses convey a semantic relation, for example the synonymy relation:

DEFINIENDUM (/PONCT parfois/ADV Apaiang/NPP ,/PONCT même/ADJ prononciation/NC)/PONCT être/V/DEF_REL un/DET atoll/NC de/P le/DET république/NC du/P+D Kiribati/NPP ./PONCT (Wikipedia)

For these, we first extract the clause as a synonymy relation for the given definiendum.

Negative adverbials cannot not be removed, as they totally change the meaning of the sentence.

Other subordinate clauses denote a domain restriction: with SDMC, we identify the most frequent cases, which derive into the following two lexico-syntactic pattern, expressed in semi regular expression:

^((?:en\|dans\|à\|sur\|selon\|pour\|chez\|par).{5,150}?)\t, \/PONCT\t/
DEFINIENDUM\t, \/PONCT\t((?:en\|dans\|à\|sur\|selon\|pour\|chez\|par).{5,150}?)\t, \/PONCT

Adverbials and subordinate clause removal obviously results in a simplification of sentences, easing the following extractions.

9 Unification of nominal phrases

Most of the time, the definiens is composed of a nominal phrase followed by complements (adjectival clauses or relative clauses). The first nominal element is therefore the hypernym of the definiendum. A series of phenomena complexify the identification of this nominal. Mainly: multiword determiners, (*a great variety of,...*) quantifiers (*three thousand ...*) and trigger words (*a kind of...*).

To overcome these cases, we rely on the tokenization process, which has recognized most of the multiword determiners, as well as trigger words, and unify only the remaining elements, based on an SDMC processing working on sequences beginning with a determiner and ending with a relative clause. We end up with three main lexico-syntactic patterns for identifying most of the nominal phrases :

N (ADJ) ? de/P N (ADJ) ?
N (ADJ){0,3}
PN+

10 Results

The linguistic preprocessing improves greatly the extraction process, as will be seen in table 1.

11 Conclusion and future work

In this contribution, we have shown through an experiment that the distributionalist hypothesis and

the accompanying computational models, can benefit from a linguistic preprocessing of corpora, especially in tasks connected to predicate-arguments structures. That derives from the fact that language has not only a sequential structure but also a hierarchical one linking lexical units to phrases, phrases to predicate-argument structures and also essential versus peripherical elements at each level. Depending on the task, any probabilistic model should preprocess the peripherical elements to eliminate noisy analysis.

References

Baroni M. and Alessandro Lenci, 2010. Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics* 36(4):673-721

Béchet N., Cellier P., Charnois T., and Crémilleux B., 2012. Discovering linguistic patterns using sequence mining. In Alexander F. Gelbukh, editor, *13th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2012*, volume 7181 of Lecture Notes in Computer Science, pages 154–165. Springer, 2012.

Blacoe W. and Mirella Lapata. 2012. A Comparison of Vector-based Representations for Semantic Composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 546–556, Jeju Island, Korea, July. Association for Computational Linguistics.

Bullinaria John A. and Joseph P. Levy. 2012. Extracting Semantic Representations from Word Co- occurrence Statistics: Stop-lists, Stemming and SVD. *Behavior Research Methods*, 44:890–907.

Clark S. 2015. Vector Space Models of Lexical Meaning (to appear). In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantics*. Wiley-Blackwell, Oxford.

Croft W. & Cruse D.A. 2004. *Cognitive Linguistics*. Cambridge UK : Cambridge University Press.

Croft, William A. 2007. Construction Grammar. In H. Cuyckens and D. Geeraerts (eds.), *The Oxford Handbook of Cognitive Linguistics*. Oxford: Oxford University Press, 463-508.

Croft,W. 2001. *Radical construction grammar: Syntactic theory in typological perspective.* Oxford: Oxford University Press.

Fillmore, Charles J., Paul Kay and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of Let alone. *Language* 64/3, 501-?-538.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*, pp. 1–32. Blackwell, Oxford.

Geeraerts, D., Grondelaers, S., & Bakema, P. 1994. *The structure of lexical variation. A descriptive framework for cognitive lexicology.* Berlin etc.: Mouton de Gruyter.

Goldberg, Adele E. 1995. *Constructions. A Construction Grammar Approach to Argument Structure.* Chicago: University of Chicago Press.

Goldberg, Adele. E. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7/5, 219–224

Harris, Z. 1954. Distributional structure. *Word*, 10(2-3):1456–1162.

Kiela, D. and Stephen Clark, "A Systematic Study of Semantic Vector Space Model Parameters," in *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, 2014, pp. 21–30

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004) The Sketch Engine. In: Williams G. and S. Vessier (eds.), *Proceedings of the XI Euralex International Congress*, July 6-10, 2004, Lorient, France, pp. 105-111.

Langacker, R. W. 1987. *Foundations of cognitive grammar. Vol. 1, Theoretical prerequisites*. Stanford, CA: Stanford University Press.

Langacker, R. W. 1991. *Foundations of cognitive grammar. Vol. 2, Descriptive application*. Stanford, CA: Stanford University Press.

Pado, Sebastian and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Schmid H.-J. 2007. Entrenchment, salience and basic levels. In: Dirk Geeraerts and Hubert Cuyckens, eds., *The Oxford Handbook of Cognitive Linguistics*, Oxford: Oxford University Press, 117-138.

Schmid H.-J. and Küchenhoff H. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24(3), 531-577.

Stefanowitsch, Anatol, and Stefan Th. Gries. 2003. Collostructions: investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8/2, 209-?-243.

Turney, Peter D. & Patrick Pantel (2010), From frequency to meaning: Vector space models of semantics, *Journal of Artificial Intelligence Research* 37:141–188.

Compiling Corpus for School Children to Support L1 Teaching: Case of Czech

Anna Čermáková
ICNC, Charles University in Prague
anna.cermakova@ff.cuni.cz

Lucie Chlumská
ICNC, Charles University in Prague
lucie.chlumska@ff.cuni.cz

1 Introduction

Corpus linguistics tools and methods have proven to be extremely successful in language (L2) teaching and there is a vast amount of corpus-based research in the field of EFL (English as foreign language). Researchers here work with e.g. general language corpora that represent the desired outcome of learning while comparing them to learner corpora i.e. authentic language produced by L2 learners. Another research avenue is concentrated around the so called developmental corpora containing language produced by children acquiring their first language. These corpora include also other languages than English (e.g. corpora included in the CHILDES project). The research based on these corpora is mostly focused on child language acquisition.

There has been, so far, considerably less attention devoted to the potential of applying corpus-based mother tongue (L1) teaching in elementary (and/or secondary) schools. Using corpora in L1 teaching makes use of highly effective exploratory learning and as Sealey and Thompson (2004: 90) suggest, corpora present a "chance to make grammar contextualized and descriptive rather than decontextualized and prescriptive". Contextualizing grammar in this sense means as well an additional stress on the exploration of lexis during the teaching process. The importance of corpora in lexicography is unquestioned. However, only recently it has been suggested that corpora used for creating dictionaries aimed at children need to be specific (Wild et al. 2013).

Mother tongue teaching in Czech elementary and secondary schools is traditionally heavily focused on competences in orthography and grammar and very little exploratory teaching is used. However, as previous work by Thompson and Sealey during their CLLIP project (Corpus-based Learning about Language in the Primary School) suggests, corpora can be successfully used already with very small children. Obviously, if integrated into the curricula, this will also require further specific teacher training (e.g. Hunston 1995, Chambers 2007). It is also clear that corpora that are available are not readily suitable as pedagogical resource (Braun 2007). This led us to a question what texts should a corpus for children contain to be useful in mother tongue teaching in schools and what other basic criteria it should meet.

2 Corpora for children

A corpus represents a source of authentic language and the issue of authenticity in language teaching has been extensively discussed (e.g. Sealey & Thompson 2007: 13; Carter 1998; Cook 1998, 2001; Stubbs 2002; Widdowson 2000). It brings forward mainly the question whether teachers should simplify the language and thus make their examples more accessible to learners, especially when small children are concerned. One of the solutions suggested by researchers is using for child learners corpora made up of language children are familiar with, that is writing for children (e.g. Thompson & Sealey 2007).

Pilot corpus work with young learners (8–10 years old) (CLLIC project) had confirmed that it was important to work with language children understood well, that is with texts they had likely read. This naturally raises questions whether writing for children (children's literature in this case) still represents authentic language or whether it is more a simplified version of the 'general language' (Thompson & Sealey 2007: 2). When discussing the language of the literature for children, we need to also consider additional features such as the ideology that is present in these texts (e.g. Hunt 1992; Sealey 2000; Wall 1991; Knowles & Malmkjær 1996). As Hunt says (1992: 2), these texts are "of massive importance educationally, intellectually, and socially" and therefore, it needs to be also assessed, how these texts influence the learning processes, including acquisition of beliefs, opinions, attitudes (van Dijk 1981). While equally important, this is a different research line.

Alison Sealey and Paul Thompson (Sealey & Thompson 2006; Thompson & Sealey 2007) had carried out a linguistic analysis of a small corpus of children's fiction (BNC subcorpus of 700 000 words) in comparison with two reference corpora (fiction for adults and newspaper texts). This comparison revealed in many respects a close similarity between the two fiction corpora (both displaying linguistic features of a narrative genre), especially when overall quantitative features were concerned (Sealey & Thompson 2006: 21). Closer lexical and semantic analysis has, unsurprisingly, revealed distinctive ways, in which the world of adults differs from that of children. Similar findings based on a much larger corpus are presented by Wild et al. (2013) in their comparison of keywords and word sketches in the Oxford Children's Corpus and

the Oxford English Corpus.

This study aims to partly replicate on the Czech language data the above mentioned research by Sealey & Thompson and Wild et al. and further explore the question what a corpus aimed at school children should look like. Based on our own "hands on" experience with school children and their teachers, we have focused on three major questions: 1) how big the corpus is to be; 2) what texts it should consist of; and 3) what functions the interface aimed at pupils and their teachers should ideally include. The third question will, for the time being, be put aside.

3 Size of the corpus

We initially explored the option of a corpus suitable for pupils from around the age 11+ (however, it may be necessary to consider a further subcorpus for older students, e.g. 15+). Based on our experience, it is clear that while the corpus should be big enough to cover the needed vocabulary, it should not be too big as both pupils and teachers find it off putting handling too big data. Big data also pose further requirements on teacher training for appropriate interpretation of the results yielded by the data.

Assessing the desired size of the vocabulary to be covered is not trivial. The Czech national curriculum for both primary and secondary schools does not have anyhow specified core vocabulary for various levels of school education, nor there are dictionaries of Czech aimed at school children or up-to-date research in this field. Czech is a highly flective language, therefore it may not be directly comparable with English but some inspiration can be drawn from English education material. We have taken as our starting point the size of the *Oxford English Dictionary for Schools* (age 11+), which contains 18,700 headwords. Research suggests that an average student acquires approximately 3,000 new words each year and an average 12th grader (age 17-18) possesses a reading vocabulary of approximately 40,000 words (Gardner 2004:1). This is, therefore, the corpus size range we shall be aiming at (for a corpus of Czech to contain 40,000 lemmas the size would have to be over 8 million words).

4 Texts in the corpus

We have decided to continue the research line suggested by Sealey and Thompson and use children's literature as the most accessible authentic material for children to explore linguistically. We have conducted our analysis of the relevant linguistic features in a subcorpus of the Czech National Corpus labelled JUN, which is broadly fiction aimed at children and young readers; all disclaimers of what constitutes children's literature do apply and need further separate investigation (see e.g. Hunt 1991; Knowles & Malmkjær 1996: 1-2; Wild et al. 2013: 193). Currently, the size of the JUN subcorpus that is available is about 4.76 million words. Since Czech is a comparatively "small" language, a large proportion of the fiction on the reading market is translated literature. The children's literature is no exception and the JUN corpus contains app. 57 % of translations. The possible translated language effect (see e.g. Puurtinen 2003; Mauranen & Kujamäki 2004) may have to be further examined.

5 Linguistic analysis of JUN corpus

We have explored the JUN corpus in terms of the overall frequency characteristics in comparison with three reference corpora: BEL (fiction for adult readers), PUB (newspaper texts), and SKRIPT (children's school essays). We have looked at the distribution of POS, the most frequent vocabulary overall and specifically compared most frequent lexical verbs, adjectives, nouns, and adverbs in the respective corpora (cf. Sealey & Thompson 2006; Thompson & Sealey 2007). We have further examined the JUN corpus in terms of keywords (with reference corpus of adult fiction) and identified some key semantic areas typical for the corpus of writing for children (cf. Wild et al. 2013).

Additional qualitative analysis was aimed at evaluative lexis and collocation profiles of some of the most frequent adjectives. Most of our findings are in line with those of Sealey and Thompson and Wild et al. Both JUN and BEL corpora, the two fiction corpora, show a significant similarity but a more detailed qualitative analysis shows considerable differences as well. SKRIPT corpus, representing student writing, is thematically fairly heterogenic, however it serves as a very useful benchmark in our comparisons.

References

Braun, S. (2007). Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. ReCALL 19(03), 307-328.

Carter, R. (1998). Orders of reality: CANCODE, communication, and culture. ELT Journal 52, 43–56.

Cook, G. (1998). The uses of reality: A reply to Ronald Carter. ELT Journal 52, 57–63.

Cook, G. (2001). 'The philosopher pulled the lower jaw of the hen'. Ludicrous invented sentences in language teaching. Applied Linguistics 22, 366–387.

Chambers, A. (2007). Popularising corpus consultation by language learners and teachers. In E. Hidalgo et al. (Eds.), Corpora in the foreign language classroom, 3 -

16. Rodopi.

Gardner, D. (2004). Vocabulary Input through Extensive Reading: A Comparison of Words Found in Children's Narrative and Expository Reading Materials. Applied Linguistics 25(1), 1–37.

Hunston, S. (1995). Grammar in teacher education: The role of a corpus. Language Awareness 4(1), 15-31.

Hunt, P. (Ed.) (1992). Literature for Children: Contemporary Criticism. London and New York: Routledge.

Knowles, M. & Malmkjær, K. (1996). Language and Control in Children's Literature. London and New York: Routledge.

Mauranen, A. & Kujamäki, P. (Eds.) (2004). Translation Universals. Do they exist? Amsterdam – Philadelphia: John Benjamins.

Oxford English Dictionary for Schools (2006). Ed by R. Allen. Oxford: OUP.

Puurtinen, T. (2003). Genre-specific Features of Translationese? Linguistic Differences between Translated and Non-translated Finnish Children's Literature. Literary and Linguistics Computing 18(4), 389–406.

Sealey, A. (2000). Childly language: children, language, and the social world. Harlow: Longman.

Sealey, A. & Thompson, P. (2007). Corpus, Concordance, Classification: Young Learners in the L1 Classroom. Language Awareness 16(3), 208–223.

Sealey, A. & Thompson, P. (2004). 'What do you call the dull words?' Primary school children using corpus-based approaches to learn about language. English in Education 38 (1), 80–91.

Stubbs, M. (2002). On text and corpus analysis: A reply to Borsley and Ingham. Lingua 112 (1), 7–11.

Thompson, P. & Sealey, A. (2007). Through children's eyes? Corpus evidence of the features of children's literature. International Journal of Corpus Lingustics 12 (1), 1–23.

Van Dijk, T. A. (1981). Discourse studies and education. Applied Linguistics, 2(1), 1–26.

Wall, B. (1991). The Narrator's Voice: the dilemma of children's fiction. London: Macmillan.

Wild, K., Kilgarriff, A., & Tugwell, D. (2013). The Oxford Children's Corpus: Using a Children's Corpus in Lexicography. *International Journal of Lexicography*, *26*(2), 190–218.

Widdowson, H. (2000). On the limitations of linguistics applied. *Applied Linguistics 21,* 3–25.

The ideological representation of benefit claimants in UK print media

Ben Clarke
University
of Portsmouth
ben.clarke@port.ac.uk

In the spirit of Halliday (1985), critical discourse analysis is said to be an ideologically committed form of social action, having, as one of its chief goals, due consideration for the interests of those afforded limited power in the social hierarchy of a community (van Dijk 1993). During times of economic and fiscal downturn it would appear valuable to ask if – and, if so, how – language and other semiotic modes (Bennett 2012) are used to maintain and reinforce the status quo.

The data for the present empirical project are all news articles published in mainstream British newspapers during the current UK government (i.e. from 12th May 2010 to the present day) and containing the expression 'benefit(s) claimant(s)'. This amounts to a dataset, held electronically as a corpus, of some three and a half thousand articles, totalling approximately 2.5 million words. Conducted in the corpus-based critical discourse analysis tradition, linguistic features which are used in the construction of benefit(s) claimants in ideologically loaded ways are identified and discussed for their significance.

Perhaps most obviously, the ideological construction at work is revealed in terms of which of the central social actors have their voices reported by the British print press (van Leeuwen 1996; Bell 1991), itself a matter, in part, of which events involving benefit claimants get reported. Typically, the voices of governmental officials and spokespersons are those which are represented, and less frequently so media protagonists too are heard; for example, with the report verb 'tell'

> Echoing advice of Norman Tebbit in 1980s, Tory minister **tells** benefit claimants to move around the country to find work Incentives for unemployed to move to get work. (The Telegraph, June 2010)

> Sources within the DWP have **told** The IoS that a realistic national roll-out - regardless of the department's public assurances - is already a year behind schedule amid fears that "technical issues over computer software" could push that back further. (The Independent, November 2012)

> Mr Osborne **told** MPs - rowdy after the Budget was leaked by a London newspaper on Twitter minutes before he delivered it - that there were no "easy answers", only more "difficult decisions" ahead. (The Daily Star, March 2013)

In contrast, the British press's mentions of benefit claimants all but exclude the verbal activity of those whose presence in the text accounts for the assembly of the dataset under study; benefit claimants voices are supressed.

When depicted in roles other than as Sayer participants in verbal activity, reference to any benefit(s) claimant(s) is disproportionally large in favour of occurrences in the plural (3,637 occurrences) rather than singular (3,941 occurrences) when compared to the equivalent for other prominent actors (e.g. MP: 2,130 occurrences – MPs: 1,226 occurrences; journalist: 182 occurrences – journalists: 114 occurrences). Such collective references serve to impersonalise (van Leeuwen 1996; Baker and McEnery 2005).

More subtle than both of the aforementioned are ideologically loaded representations deriving from the use of adjectives which have apparent connotative meanings. 'Tough', for example, is used – often in conjunction with the first of the linguistic patterns mentioned above – in a way which taps into familial discourses as revealed in its collocational association with 'love' in the present corpus.

> […] a strong sanctions regime was put in place which allowed case managers to have a "**tough love**" relationship with claimants. If you didn't play ball, you didn't get benefits. (The Telegraph, September 2010)

> […] unemployment FELL by 46,000 to 2.56million in the three months to July 1. It was the lowest level for a year - and Mr Duncan Smith said it proved the Government's "**tough love**" policies to get people off handouts were working. (The Sun, August 2012)

> Labour, which will vote against the measure, will try today to answer Tory claims that it is "soft" on scroungers by announcing a "**tough love**" plan to force adults who have been out of work for more than two years to take up a government "job guarantee" or lose their benefits. (i-Independent, January 2013)

These familial discourses – reinforcing, as they do, the hierarchy of social structure – are also carried forward by the tendency for benefit(s) claimants to serve in participant roles which undergo actions of verbal, cognitive and physical sorts (Halliday 1994) with government and government-related referents typically initiating such actions; for example:

> And the Government [Sayer] will **encourage** [verbal process] honest benefit claimants [Receiver] to shop the cheats to a DWP hotline [Verbiage]. (The Daily Mail, December 2010)

> I FULLY support the Government's initiative to **expect** […mental process] benefits claimants to do some work for the money they receive […Phenomenon] (The Express, January 2012)

> […] it emerged that Labour-run Newham council [Actor] was planning to **move** [material process] housing benefit claimants [Goal] to Stoke-on-Trent. (The Guardian, April 2012)

Yet when initiating, rather than undergoing, action, benefit(s) claimants are typically cast as dependent Sensors of cognitive activity; for example:

> Even in sparsely populated Cornwall benefit claimants [Sensor] **need** [mental process] £5m to cover their losses [Phenomenon]. (The Guardian, November 2010)

> The claimants [Sensor] **want** [mental process] the judges to grant them income support of GBP 65.45 a week or pension payments if they are old enough [Phenomenon]. (The Guardian, April 2012)

These and further linguistic trends like them reveal the media's role in popularising a particular view of benefit claimants. As per Fairclough (1995), van Dijk (1993) and colleagues, the approach adopted here is that of identifying those linguistic strategies which contribute to the creation of prevalent discourses of the type here discussed; this is done in order to de-naturalise such discourses, itself a first step in re-addressing the balance and providing space for alternative and counter discourses.

References

Baker, P. and McEnery, T. 2005. "A corpus-based approach to discourses of refugees and asylum seekers

in UN and newspaper texts". *Journal of Language and Politics* 4 (2), 197-226.

Bell, A. 1991. *The language of news media*. London: Wiley-Blackwell.

Bennett, J. 2012. "Chav-spotting in Britain: the representation of social class as private choice". *Social Semiotics* 23 (1): 146-162.

Fairclough, N. 1995. *Critical discourse analysis: The critical study of language*. London: Longman.

Halliday, M.A.K. 1985. 'Systemic Background'. In J.D. Benson and W.S. Greaves (eds.) *Systemic perspectives on discourse*. Norwood, New Jersey: Ablex, 1-15.

Halliday, M.A.K. 1994. *An introduction to functional grammar*. 2nd edition. London: Edward Arnold.

van Dijk, T.A. 1993. *Principles of critical discourse analysis*. *Discourse and Society* 4: 249-283.

van Leeuwen, T. 1996. "The representation of social actors". In C.R. Caldas-Coulthard and M. Coulthard (eds.) *Text and practices: Readings in critical discourse analysis*. London: Routledge: 32-70.

"I crave the indulgence of a discriminating public to a Work": Effective interaction between female authors and their readership in Late Modern scientific prefaces and works

Begoña Crespo
Universidade da Coruña

Much of what we nowadays term "front matter" in certain kinds of publications was conceived of in the past as a direct address to the reader under different labels. Therefore, prefaces, forewords, dedications and addresses to the reader of all sorts were basically placed at the beginning of literary and non-literary works with the clear intention of attracting the attention of the audience. In the course of time, standard formulae from classical tradition were developed and certain rhetorical devices consolidated. In a similar vein, there was an evolution of the way in which scientific knowledge was transmitted and also an evolution of the style considered more appropriate for such an objective. However, authors were familiar with the patterns of prefaces and dedications which were highly conventionalised, a fact that contrasted with the so to speak "discursive freedom" they were allowed when composing their scientific works.

My main research question in this piece of work is whether prefaces to scientific works and the body of the texts themselves show the same the degree of involvement or detachment or not and in what sense. To this end, I will analyse scientific texts and their corresponding "front matter" written by women between 1700 and 1900. All samples will be extracted from different sub-corpora of the *Coruña Corpus of English Scientific Writing*, namely, *CETA* (Corpus of English Texts on Astronomy, 2012), *CEPhiT* (Corpus of English Philosophy Texts), *CELiST* (Corpus of English Life Sciences Texts) and *CHET* (Corpus of Historical English Texts). The Penn-Helsinki and the *Corpus of Historical American English* (*COHA*) will be used as reference corpora when possible, especially to look into the use of certain linguistic strategies (see below) in the body of scientific works and other works in general from the same period. Although the number of samples written by female authors is not very high, this should not be an obstacle for my main purpose here, that is to say, the comparison of how these authors use linguistic features denoting involvement in their works and their prefaces since the scarcity of female scientific writings is nothing but a mirror of eighteenth- and nineteenth- century reality.

I will focus on the use of some of the linguistic

elements generally admitted to express or denote involvement and interaction (Biber, 1988; Prelli, 1989; Lakoff, 1990; Besnier, 1994). These features include the use of first and second person pronouns, wh-questions, hedges, amplifiers and private verbs. Although it is not a working hypothesis in itself, it can be expected to find more of these features in prefaces than in other text-types as they may have been used as a strategy to contact the members of the epistemic community more directly (Narrog, 2012). This might be so since one of the primary pragmatic functions of prefaces and front matter in general is to exert a positive influence on the readership (Bradbury-Jones et al., 2007). With this study we will have the opportunity to prove whether this hypothesis is true in all text-types under survey.
The analysis of the above features in relation to the two variables mentioned (time and text-type) will offer a glimpse of language change and variation in scientific discourse. It will hopefully provide as well a preliminary portrait of the evolution towards detachment in specialised registers as observed in the twentieth century in the hands of writers who are classically considered more involved: female writers (Argamon et al., 2003).

References

Argamon, Shlomo; Koppel, Moshe; Fine, Jonathan; Shimoni, Anat. 2003. Gender, Genre, and Writing Style in Formal Written Texts. *Text* , 23/3.

Besnier, Niko. 1994. Involvement in linguistic practice: An Ethnographic Appraisal. *Journal of Pragmatics* 22: 279-299.

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge, UK: Cambridge University Press.

Bradbury-Jones, Caroline; Irvine, Fiona; Sambrook, Sally. 2007. Unity and Detachment: A Discourse Analysis of Doctoral Supervision. International *Journal of Qualitative Methods*: 81-96.

Lakoff, Robin T., 1990. *Talking power: The politics of language in our lives*. New York: Basic Books.

Narrog, Heiko. 2012, *Modality, Subjectivity, and Semantic Change: A Cross-Linguistic Perspective*. Oxford: Oxford University Press.

Prelli, Lawrence J. 1989. The rhetorical construction of scientific ethos. In: Herbert W. Simon (ed.), *Rhetoric in the human sciences*. London: Sage.

Using corpora in the field of Augmentative and Alternative Communication (AAC) to provide visual representations of vocabulary use by non-speaking individuals

Russell Thomas Cross
Prentke Romich Company
rtc@prentrom.com

1 Introduction

The field of Augmentative and Alternative Communication (AAC) is a discipline focuses on the needs of individuals with severe speech or language problems, the severity of which is such that there is a need to supplement any existing speech or replace speech altogether.

Aided communication methods include low-tech solutions such as paper and pencil to communication books or boards populated by words and/or symbols, or devices that produce voice output (speech generating devices or SGDs) along with text output. Electronic communication aids allow the user to use picture symbols, letters, and/or words and phrases to create messages. Some devices can be programmed to produce different spoken languages.

The success individuals may have in using an SGD is heavily influenced by the amount of time spent by parents and spouses, educators, Speech and Language Therapists, in helping them to learn how to use the system (Arnott & Alm, 2013; Ball & Lasker, 2013; Travis & Geiger, 2010).

2 Improving performance using automated data logging (ADL)

Automatic data logging is a feature of some voice output communication aids. Such data can be useful in providing clinicians with information on how a client is using a device and, more importantly, how well that client is using it to communicate effectively. There are limitations to the data, which include;

- Absence of input from communication partners
- Absence of any multi-modal elements.
- Absence of social/geographical context.
- Need to mark explicitly if someone else is using the device for modeling/teaching.

Given that these limitations are recognized, it is still possible to use the information in a fruitful and constructive way. For example, one simple measure of AAC use is to count words used, which can give an idea of an individual's knowledge of the lexicon

available to them in their AAC system. Another is to measure the time period between linguistic events so as to get an idea of communication rate. A third is to look at the type of words being used and determine the spread of different parts of speech.

3 Visualizing the data

One challenge with machine-logged data is that in its raw form it can be difficult to interpret. It is possible to use manual and semi-automated systems such as SALT (Miller & Chapman, 1983) AQUA (Lesher, Moulton, Rinkus, & Higginbotham, 2000), PERT (Romich, Hill, Seagull, Ahmad, Strecker, & Gotla, 2003) and QUAD (Cross, 2010) to convert such raw data into more user-friendly formats. Another method is to use specific data visualization software that is designed to convert numeric and textual data into graphic formats.

Cross (2013) developed a web-based automated data analysis software that allows for the uploading of a log file to a secure server, where it can be parsed in a number of ways to as to present summary data in the form of a visual dashboard. The current version allows for data to be analyzed in terms of;

- Word frequency
- Parts of Speech
- Performance against target vocabulary
- Daily/Weekly/Monthly device use

It's also possible to search for specific instances of words and see them in context.

4 Using the Corpus of Contemporary American English

To provide a large corpus against which client-generated utterance could be matched, the Corpus of Contemporary America English (Davies, 2008) was used. This was chosen because not only did it provide a very large database – far larger than any currently available in the field of AAC – but it also includes frequency data and grammatical tagging based on the CLAWS system (Garside, 1987). Both word frequency and syntax (mainly in the area of morphology) are important pieces of information when monitoring the performance of an aided communicator (Binger, 2008; Binger & Light, 2008). Furthermore, such information can inform educational and clinical intervention programs (Cross, 2013).

Another feature of the database is that words are lemmatized, providing a level of analysis that has implications for the teaching vocabulary as *word sets* rather than individual lexical items. For example, if a client demonstrates the use of *jump*, *jumps*, *jumped*, *walks*, and *walking*, teaching *jumping* and *walked* to "complete the set" makes sense.

5 Outline of how the system works

The basic operation of the server is fairly simple. It consists of three elements:

(a) **Uploaded Data File**: The primary input to the system is a plain text (TXT) file that has been created by the automated data logging feature of an SGD.

All individual uploads are aggregated over time and become the basis of a "merged file" that provides a personal database of language use. It is this aggregated database that is used for all the different types of analyses the system has to offer.

(b) **Comparison Database**: Certain analyses – such as the "Parts-of-Speech" analysis, use the database in order to identify and present words. The system makes use of color coding in order to represent these in order to create, for example, a bar chart:

(c) **Analysis "widgets":** Specific analyses can be performed by selecting a "widget" - a single-hit button that triggers a particular action. For example, a "Cloud" widget looks at all the words used in the merged file within a specific time period and then displays these as a word cloud picture, where the size of a word is directly proportional to its frequency of use.

As another example, a "Weekly Use" widget counts the number of times within a 15-minute period that the SGD is used. It then displays this as a graph.

The graphical results of using any of these widgets can be saved as PNG graphics files and then used to create reports and summaries.

6 Next Steps

Using client-generated data to improve the performance of individuals who use SGDs is still relatively new. The use of large scale corpora to provide enable comparisons to be made and individual performance to be tracked is also in its infancy. This means that the metrics being used are rather broad and need to be made more granular and specific. For example, the analysis of parts-of-speech uses the global categories of noun, verb, adjective etc. but a more precise breakdown using specific CLAWS tags would yield much more information.

Another challenge is to be able to use more flexible filters in the system so as to be able to break down the data into more focused conditions. Being able to have the server handle questions such as "how many times was the *–ing* participle used one month ago compared with this week" is

pedagogically value.

References

Arnott, J. L., & Alm, N. (2013). Towards the improvement of Augmentative and Alternative Communication through the modelling of conversation. *Computer Speech & Language*, 27(6), 1194-1211.

Ball, L. J., & Lasker, J. (2013). Teaching Partners to Support Communication for Adults with Acquired Communication Impairment. Perspectives on *Augmentative and Alternative Communication*, 22(1), 4-15.

Binger, C. (2008). Grammatical Morpheme Intervention Issues for Students Who Use AAC. Perspectives on *Augmentative and Alternative Communication*, 17(2), 62-68.

Binger, C., & Light, J. (2008). The morphology and syntax of individuals who use AAC: research review and implications for effective practice. *Augmentative and Alternative Communication*, 24(2), 123-138.

Cross, R. T. (2010). Developing Evidence-Based Clinical Resources Embedding. In Hazel Roddam and Jemma Skeat *Evidence-Based Practice in Speech and Language Therapy* (pp. 114-121): John Wiley & Sons, Ltd.

Cross, R. T. (2012). Using AAC device-generated data to develop therapy sessions. Paper presented at the *American Speech Hearing and Language Association Annual Convention*, Atlanta, GA.

Cross, R. T. (2013). The Value and Limits of Automated Data Logging and Analysis in AAC Devices. Paper presented at the *ASHA Convention*, Chicago, IL.

Davies, M. (2008-). *The Corpus of Contemporary American English: 425 million words, 1990-present*. Available online at http://corpus.byu.edu/coca

Garside, R. (1987). The CLAWS Word-tagging System. In R. Garside, G. Leech & G. Sampson (Eds.), *The Computational Analysis of English: A Corpus-based Approach* (pp. 30-41). London: Longman.

Lesher, G. W., Moulton, B. J., Rinkus, G., & Higginbotham, D. J. (2000). A Universal Logging Format for Augmentative Communication. Paper presented at the *2000 CSUN Conference*, Los Angeles. http://www.csun.edu/cod/conf/2000/proceedings/0088Lesher.htm

Miller, J., & Chapman, R. (1983). *SALT: Systematic Analysis of Language Transcripts*. San Diego: College Hills Press.

Romich, B. A., Hill, K. J., Seagull, A., Ahmad, N., Strecker, J., & Gotla, K. (2003). AAC Performance Report Tool: PERT. Paper presented at the *Rehabilitation Engineering Society of North America (RESNA) 2003 Annual Conference*, Arlington, VA.

Travis, J., & Geiger, M. (2010). The effectiveness of the Picture Exchange System (PECS) for children with autism spectrum disorder (ASD): A South African pilot study. *Child Language Teaching and Therapy*, 26(1), 39-59.

Changing Climates: a cross-country comparative analysis of discourses around climate change in the news media

Carmen Dayrell
Lancaster University
c.dayrell@lancaster.ac.uk

John Urry
Lancaster University
j.urry@lancaster.ac.uk

Marcus Müller
University of Heidelberg
marcus.mueller@gs.uni-heidelberg.de

Maria Cristina Caimotto
University of Turin
mariacristina.caimotto@unito.it

Tony McEnery
Lancaster University
a.mcenery@lancaster.ac.uk

1 Introduction

Although climate change has reached a broad scientific consensus with respect to its impacts and the urgent need to take actions, global cooperation for its solution has not yet been achieved. There are still those who remain sceptical and challenge the scientific treatment of climate change, or at least part of it. As a result, societies worldwide differ in their level of concern and governments have taken different positions and pursued different policies.

This paper focuses on the public-sphere debate around climate change issues in Brazil and Britain. They are among the largest economies in the world and are also major emitters of greenhouse gases. At the same time, they have both adopted significant measures to curb emissions and are major players in international debates on global warming, but differ in relation to key points. Britain strongly depends on fossil fuel combustion and is among the nations with highest records of historical emissions per capita. Brazil by contrast is an emerging economy whose fossil fuel-based emissions are low by global standards due to high investment in hydropower and biofuel. Brazil has the largest reserve of agricultural land in the world and agriculture, land-use and deforestation are leading sources of its greenhouse-gas emissions. It also houses most of the Amazon forest and river basin and is expected to play a key role in the world's management of natural resources.

The level of concern about climate change is strikingly different. Recent survey shows that Brazil is a leading country in terms of concern about climate change, with nine-in-ten Brazilians considering global warming a *very* serious problem (PEW 2010). Britain show a high percentage of climate scepticism, where about one-quarter state either that climate change is *not too serious* or that it is *not a problem* at all.

This paper aims to carry out a comparative analysis of the discourses around climate change within the news media across Brazil and Britain. Our primary purpose is to investigate what kind of debate surrounding climate change issues is found within the public sphere in each society. More specifically, we aim to examine how climate change has been framed in printed newspapers in the past decade. Here, we are interested in gaining a better understanding of the role of the mass media in shaping public opinion.

1 Corpora and methods

Both the British and the Brazilian corpora on Climate Change consist of newspaper articles making reference to climate change/global warming and published by daily newspapers with wide circulation between Jan/2003 and Dec/2013. These texts were selected on the basis of a set of query words/phrases, established according to Gabrielatos (2007). The corpora include news reports, editorials, opinions, articles, and interviews. Texts in the British corpus come from seven broadsheet papers and six tabloids, both weekday and Sunday publications, totalling over 79 million words. The Brazilian component consists of 19,135 texts (10.8 million words) collected from 12 daily broadsheet papers.

This study aims to identify similarities and differences across the corpora regarding the discourses around climate change within the news media. To this end we make use of keyword analysis so as to identify salient lexical items within each year. We then resort to collocation analyses to uncover the contexts within which such keywords occur.

2 Results

Overall Brazilian mainstream media adopted, organised and mobilised a 'gradualist' discourse (Urry 2011), as best represented by the Intergovernmental Panel on Climate Change (IPCC) Reports. Much media debate in Brazil is organised explicitly or implicitly around how to deal with the 'reality' of climate change while climate scepticism was almost non-existent.

Salient lexical itens throughout this period include: IPCC, UN Framework Convention on Climate Change (UNFCCC), UN Conference on Climate Change, Brazilian Forum on Climate Change, and Kyoto Protocol. The gradualist discourse is also evident through collocation analyses of keywords such as *climate, change,*

global, warming, *deforestation, greenhouse, gases, carbon, emissions, energy, temperature, fuel, oil, fossil,* and *coal*. For example, significant collocates of *climate change* include: *impact(s), consequence(s), effects* as well as *action(s), combat, fight, face, avoid, mitigation,* and *adaptation*. These lead to a explicit discussion on the consequences of climate change and the urgent need to take actions.

Although the gradualist discourse is also voiced within the British media, it is not as explicit as in the Brazilian media. For example, the *IPCC* and the *Kyoto Protocol* are mentioned four times more frequently in the Brazilian than in the British media: 8.6 and 6.9 mentions per 100,000 words respectively in British newspapers in relation to 31.6 and 35.7 mentions in Brazilian newspapers.

In relation to the Brazilian media, the British media gives much more room to climate change scepticism. The collocations of *climate change* illustrate well how Brazil and Britain differ from each other in terms of discussions around climate change issues. Highly significant within British newspapers are the collocations of climate change with *questions, denier(s), sceptic(s), denial,* and *scepticism*, which mostly refer to voices of climate-change scepticism. Here is an example: *A preliminary study of 6,000 logbooks has produced results that raise* ***questions*** *about* ***climate change*** *theories* (The Sunday Times, 03/Aug/2008).

At the same time, we also find various collocations of *climate change* that are related to the gradualist discourse, for example: *combating, mitigating, mitigate, manmade, combat, adapting, induced, mitigation, irreversible, tackling, posed, adapt, impacts, poses, addressing, avert, adaptation,* and *tackled*. Interestingly, *cost* figures as the most salient collocate of *climate change* within the British media, referring to discussion around the human, social, and economic costs of dealing with climate change. Such debate is not salient within the Brazilian media.

3 Final Remarks

This study is currently being extended to Germany and Italy. Like Brazil and Britain, these are also major emitters of greenhouse gases. However, Germany has efficient public transport with a highly organised structure for biking. Renewable energy rates are high and can reach 75% of domestic and industrial energy use on certain days. Italy on the other hand is the country of 'Ecomafia' where corruption is directly related to illegal activities that harm the environment.

As regards public opinion, Germany seems to stand somewhere between Brazil and Britain. Although climate scepticism was higher (14%) than in Brazil (3%) (PEW 2010), some studies have indicated that Germans are fairly sensitive to the environmental risks of different technologies. The level of seriousness Italians attribute to the problem of climate change is similar to that of Britain (Italy 42%, UK 44%, Germany 66%) but Italians are among the least likely in Europe to express personal responsibility for climate action (Italy 5%, UK 20%, Germany 36%) (EC 2011).

The German and Italian corpora are currently being built, according to the same criteria used to compile the Brazilian and British corpora. The German corpus is expected to contain approximately 40 million words from five broadsheet papers and five magazines. The Italian corpus is estimated to contain about 10 million words from nine major Italian newspapers.

Thus, this paper will discuss relevant aspects of the discourses around climate change issues within the news media of four relevant countries: Brazil, Britain, Germany and Italy. Such analysis can provide useful insights and enhance our understanding of how society and media coverage interact within the climate change context. This is important because society is central to high carbon lives; moving from a high- to a low-energy economy involves changing social practices (Urry 2011).

4 Acknowledgements

This research is part of the Changing Climates project currently being conducted at the ESRC-funded Centre for Corpus Approaches to Social Science (CASS), Lancaster University (grant reference: ES/K002155/1).

References

EC (European Commission). 2011. *Climate Change*. *Special Eurobarometer 372 Report*. Available at: http://ec.europa.eu/public_opinion/archives/ebs/ebs_372_en.

Gabrielatos, C. 2007. "Selecting query terms to build a specialised corpus from a restricted-access database". *ICAME Journal* 31: 5-43.

PEW. 2010. *2010 Pew Global Attitudes Report. Obama more popular abroad than at home, global image of U.S. continues to benefit. Muslim disappointment*. Available at: http://www.pewglobal.org/2010/06/17/obama-more-popular-abroad-than-at-home/

Urry, J. 2011. *Climate Change and Society*. Cambridge: Polity Press.

The politics of *please* in British and American English: a corpus pragmatics approach

Rachele De Felice
University College London
r.defelice@ucl.ac.uk

M. Lynne Murphy
University of Sussex
m.l.murphy@sussex.ac.uk

Please is '[t]he most obvious example of a politeness marker in English' (Watts 2003:183). Nevertheless, little work to date has explored this politeness marker across Englishes, despite various indications that its use differs in British and American Englishes. For example, Algeo (2006) found that the interjection *please* occurs twice as frequently in the British spoken as in the American spoken portions of the Cambridge International Corpus. This paper presents a study of the use of *please* in these two varieties of English, combining insights from comparative and corpus pragmatics.

Research in comparative pragmatics is intrinsically interesting and has important applications because of the inherent dangers of impoliteness and face-threatening acts that can arise from differing conventions across cultures. Traditionally, much comparative pragmatic research has started from the level of speech act (e.g. CCSARP; Blum-Kulka and Olshtain 1984) and has asked how particular speech acts are realised in the languages of different cultures. Marked differences in such realisations can be found even within "inner-circle" Englishes, as has been demonstrated through corpus investigations of, for example, suggestions (Flöck 2011) and conversation openings (Schneider 2012).

Another way to approach comparative pragmatics is at the lexical level, where use and function of pragmatic markers can be compared (e.g. Fung and Carter 2007; several of the papers in Romero-Trillo 2008; Aijmer 2013). This is a natural approach in corpus linguistics, and one that has become possible as larger corpora of different learner and non-learner language varieties have become available. While this approach allows us to discover a great deal about how particular pragmatic markers are used when they are used, searching for lexical or phrasal items can only tell us about where they occur, and not where they could have occurred, but didn't.

This problem can be overcome, and further insights gained, using well-matched, speech-act-tagged corpora (De Felice et al. 2013). Having speech-act information for each utterance enables the researcher to query the corpus on the basis of categories rather than items and to observe the pragmatic, lexical, and phraseological characteristics of a category as a whole. Our work exemplifies this corpus-pragmatics approach by using two email corpora to investigate *please* as a politeness marker in British and American English.

Watts (2003:80) identifies *please* as an 'Expression of Procedural Meaning' (EPM) that is part of 'politic' behaviour: 'when [EPMs] are missing, their absence is easily interpretable as impoliteness$_1$ and when they are in excess of what is required by the situation, they are easily interpreted as politeness'. Leech (2014: 161) describes *please* as marking an utterance 'as a request spoken with a certain (often routine) degree of politeness'. However, some informal claims (Trawick-Smith 2012; Murphy 2012) have been made that the presence of *please* in requests can seem less polite in American English than the equivalent request without it, emphasizing social power differences and expressing impatience. This raises the possibility that use of *please* is 'politic' behaviour in BrE in a way that it is not in AmE, thus creating more pragmatically marked expressions in AmE.

These hypothesized differences can be tested both in terms of whether/how often *please* occurs in requests and offers, and in terms of its position and collocates where it occurs. Sato (2008) argues that different types of facework are performed depending on the position of *please* in direct and indirect requests in American English, and in less well-matched corpora Murphy (2015) argues that these positions have different significance in BrE and AmE, as demonstrated by their proportional frequency and the modal verbs that co-occur with *please* in these positions. In BrE, Wichmann (2004) concludes that *please* 'only occurs in situations where the imposition is either minimal or socially sanctioned', that is, 'only when there is very little face-work to be done' (Wichmann 2004:1544).

Our study of *please* develops this work in a more methodologically rigorous way, by comparing both presence and absence of *please*, its position and its collocates in direct and indirect requests in two comparable business email corpora: Enron (AmE; Styler 2011) and COBEC (BrE; Anke et al. 2013; De Felice and Moreton 2014).

A randomly selected sample of 500 instances of direct and indirect requests is extracted from each corpus. As noted above, the presence or absence of *please* is correlated with its position in the utterance, its collocates, and other lexico-syntactic information such as the use of imperatives, modal verbs, or past tense. We also examine whether there are any patterns relating the extent of the imposition of the request (where this is recoverable from context) to the use or non-use of *please*.

An initial overview of the data confirms a difference in usage: *please* appears in around 50% of requests in the BrE corpus, but in only around 37% of AmE ones. However, direct and indirect requests differ, with fewer instances of *please* in AmE indirect requests compared to their BrE equivalents, but about the same in direct requests.

For example, *please* is used in almost half of BrE indirect requests of the form *can/could you* (such as *could you please fax this text to Mark?*), but in only about a quarter of the equivalent AmE instances. In particular, it appears from the data that in BrE there is a tendency to use *please* even with low imposition routine workplace requests such as sending something: *can/could you send* occurs almost always with *please* in the BrE data, and almost never with it in the AmE data (cf. BrE *can you please send me a draft copy of the document* vs. AmE *can you send the email around?*). However, this does not mean that AmE speakers do not perceive the need to moderate the force of their requests: where *please* is absent, we find other mitigators in the form of clauses such as *if possible could you…* or *when you get a chance*.

With regard to direct requests, as noted above, *please* is found equally in AmE and BrE data, occurring in around 40% of instances in both corpora. While there is some overlap in usage, with typical (and semi-fossilized) workplace expressions such as *please feel free to…*, *attached please find*, or *please forward this* common in both corpora, closer analysis of a wider range of utterances reveals interesting differences. For example, many of the AmE instances of *please* co-occur with imperatives which require concrete actions and suggest instructions being given to someone of lower status (though this is difficult to prove conclusively in the absence of the relevant social information): *please book me an additional ticket*; *please add to my calendar*; *please process the following changes*. In BrE requests, most of the imperatives describe cognitive or communicative processes which often lead to benefits for both hearer and speaker: *please give me a call*, *let me know*, *note*, *remember*. By combining corpus linguistic methods and lexical analysis, we can obtain a clearer picture of the contexts in which AmE and BrE speakers choose to use *please*, and we can empirically test the assertion that it is an unmarked expression of procedural meaning in BrE and a more marked request marker in AmE, associated with urgency and social-power differentiation.

Our study establishes whether the perceived differences between the two varieties are actually encountered in workplace communication, and will help delineate models of politeness for the two contexts. These are considered with reference to stereotypical characterisations of American culture as a positive-face-oriented solidarity system and mainstream British culture as a mixed weak-deference system (Scollon and Scollon 1983).

References

Aijmer, K. 2013. *Understanding pragmatic markers: a variational approach*. Edinburgh: Edinburgh University Press.

Anke, L., Camacho Collados, J. and Moreton, E. 2013. The development of COBEC: the Corpus of Business English Correspondence. Paper presented at the V Congreso Internacional de Lingüística de Corpus (CILC), Alicante.

Blum-Kulka, S. and Olshtain, E. 1984. Requests and apologies: a cross-cultural study of speech act realization patterns (CCSARP). *Applied Linguistics* 5(3): 196-213.

Brown, P. and Levinson, S. 1987. *Politeness: some universals in language usage.* Cambridge: Cambridge University Press.

De Felice, R., Darby, J., Fisher, A. and Peplow, D. 2013. A classification scheme for annotating speech acts in a business email corpus. *ICAME Journal* 37: 71-105.

De Felice, R. and Moreton, E. 2014. The pragmatics of Business English: introducing the Corpus of Business English Correspondence (COBEC). Paper presented at the 7th IVACS Conference, Newcastle.

Flöck, I. 2011. 'Don't tell a great man what to do': Directive speech acts in American and British English conversations. Poster presented at 12th International Pragmatics Conference, Manchester, July.

Fung, L. and Carter, R. 2007. Discourse markers and spoken English: native and learner use in pedagogic settings. *Applied Linguistics* 28(3): 410-439.

Leech, G. 2014. *The pragmatics of politeness*. Oxford: Oxford University Press.

Murphy, M. L. 2012, 18 August. Saying *please* in restaurants. *Separated by a Common Language* (blog). http://separatedbyacommonlanguage.blogspot.co.uk/2012/08/saying-please-in-restaurants.html (3 Dec 2014)

Murphy, M. L. 2015. Separated by a common politeness marker: the case of *please.* Paper submitted to International Pragmatics Association conference, July, Antwerp.

Romero-Trillo, J. 2008. *Pragmatics and corpus linguistics: a mutualistic entente*. Berlin: Mouton de Gruyter.

Trawick-Smith, B. 2012, 13 May. Impolite 'please'. *Dialect Blog*. http://dialectblog.com/2012/05/13/impolite-please/ (3 Dec 2014)

Schneider, K. 2012. Appropriate behavior across varieties of English. *Journal of Pragmatics* 44: 1022–37.

Scollon, R. and Wong Scollon, S. 1995. *Intercultural Communication*. Oxford: Blackwell.

Styler, W. 2011. The EnronSent Corpus. Boulder, CO: University of Colorado.

Watts, R. 2003. *Politeness.* Cambridge: Cambridge University Press.

Wichmann, A. 2004. The intonation of please-requests: a corpus-based study. *Journal of Pragmatics* 36: 1521–49

Collecting the new Spoken BNC2014 – overview of methodology

Claire Dembry
Cambridge University Press
cdembry@cambridge.org

Robbie Love
Lancaster University
r.m.love@lancaster.ac.uk

1 Introduction

Cambridge University Press (Cambridge) and The Centre for Corpus Approaches to Social Science at Lancaster University (CASS) are working together on a joint project to collect a new corpus of spoken British English from the mid-2010s - the *Spoken British National Corpus 2014 (Spoken BNC2014)*. This will be the first freely available corpus of its kind since the spoken component of the original British National Corpus (Leech 1993) (henceforth *Spoken BNC1994*).

This paper describes the methodology employed to collect recordings from project participants, and outlines the success and difficulties we have experienced in using this method since the project began. It should be noted that the Spoken BNC1994 consisted of two parts – *demographic* recordings of spontaneous natural conversations and *context-governed* recordings made at specific types of meetings and events, (see e.g. Aston & Burnard (1997:31) for a further discussion of these). The data collection and methodology outlined here relate only to the collection of *demographic* data.

2 Pilot study

As a precursor to the joint project Spoken BNC2014 with CASS, Cambridge ran an initial pilot study to test three factors that were key in determining if (and how) large-scale spoken data collection might be possible[28]. These were:

- a) How could we recruit and motivate a wide enough range of individuals from around the UK to collect recordings?
- b) How would contributors actually make recordings of a good enough quality to include in the corpus (i.e. for audio transcribers to easily work with)?
- c) What would be the most effective way to manage the collection of the supporting information from the data contributors (namely, speaker and recording information

[28] It should be noted that Cambridge's pilot study was separate from but related to a subsequent pilot study carried out by Lancaster (Love, 2014).

sheets and consent forms relating to all speakers in the recordings)?

As an initial test, a basic job advert was put together detailing a freelance opportunity for participants to make recordings of everyday conversations conducted with their friends and family. Participants were asked to use whatever recording device they might have available. We deliberately did not specify the type of recording device, software or set up that participants might use, so we could gauge both what technology was freely available and known about by participants, and what difference (if any) this choice made to recording quality. We also did not target or aim at any particular demographic. We offered participants a small fee (£18 per hour of recording, pro rata) in return for their recording and their completed consent forms and related information. This advert initially was made available on Cambridge's recruitment website.

Contributors were managed by a nominated project manager, who fielded all questions, sent out and received back all paperwork and maintained all databases and processes. All contact with participants (except for the signing of agreements and consent forms, which were posted out as a hard copy) was conducted entirely by email. As a result of this process, we were able to test and develop supporting documents, guidelines and FAQs that anticipate and answer common.

On the whole, participants used recording software pre-installed on their smartphone (and less so, on their tablet or laptop) to make recordings, and transferred these to Cambridge using the file hosting service Dropbox.

Pilot testing revealed that not only is it possible to encourage participants to engage with the project, but also that almost every recording we received was of a suitable quality to transcribe. Those that were excluded were largely due to the choice of recording context (e.g. excessive background noise), and not due to any problem with the recording equipment per se.

3 Scaling up

Once the core method was established, the initial aim of our joint project was to attract more contributors. In order to do this, we promoted the project in the following locations:

- Articles and features in national print and broadcast media (e.g. The Times, The Daily Mail, The Metro, BBC Radio, Sky News)
- Public talks and events (e.g. at the ESRC Festival of Social Science and the Cambridge Festival of Ideas)
- Promotion by project team members (e.g. through mentions on Twitter and Facebook, blog articles, project websites).
- Adverts on Facebook and Twitter.

Early indications suggest that Twitter advertising in particular achieves the most success for the least expense (both with respect to time and money), with around 55 project signups generated by a one week Twitter ad campaign alone.

Although our campaigns in the traditional print and broadcast media did garner good general coverage for the project, it resulted in only a handful of project signups, and so is not a useful way to find new participants.

4 Where we are and what's next

As with the Spoken BNC1994, the Spoken BNC2014 gathers demographic speaker information, namely: age; gender; accent/dialect, place of birth, location living, highest qualification and occupation. We also collect information about the topic of the recording, (e.g. computer games, DIY, family) and about the recording context (e.g. at the pub, cooking dinner).

Our efforts in publicising the project so far have attracted 175 project participants (i.e. those making recordings and completing paperwork) and 308 unique speakers. In contrast, the demographic element of the spoken BNC1994 employed 124 participants who recorded their daily interactions wearing a walkman tape recorder over a 2-7 day period (see e.g. Crowdy, 1995). The Spoken BNC2014 instead intends to reach a wider range of data contributors, who are likely to each contribute a smaller total number of words to the corpus, as compared to those who made Spoken BNC1994 recordings.

This expansion phase of the project is ongoing, but early analysis of the 240 hours of recordings collected so far shows that we have achieved a good initial range of recordings across age and gender categories. However, some underrepresentation of older (60+) and younger (under 18) speakers, along with those from certain geographical areas (in particular, Scotland, Wales and south west England) exist at this early stage. Planned strategies to tackle these problems include the introduction of a 'recommend a friend' scheme (where possible) in under-represented categories, and region-specific advertising, e.g. on Facebook and Twitter.

5 Conclusion

Both Cambridge's initial pilot study and the subsequent joint project have clearly shown that the prevalence of mobile phones and technology in our everyday lives and our familiarity with social media have meant that it is entirely possible to collect quality recordings on a large scale by commissioning members of the public as independent freelancers to make unsupervised recordings.

References

Aston, G. & Burnard, L. 1997. The BNC handbook. Exploring the BNC with SARA. Edinburgh: Edinburgh University Press.

Crowdy, S. 1995. The BNC spoken corpus, in G. Leech, G. Myers and J. Thomas (eds.) 1995, Spoken English on Computer: Transcription, Mark-up and Application. London: Longman, pp.224-235.

Leech, G. (1993). 100 million words of English. *English Today*, 9-15. doi:10.1017/S0266078400006854

Love, R. (2014). *Methodological issues in the compilation of spoken corpora: the Spoken BNC2014 pilot study*. Lancaster University: unpublished Masters dissertation.

"Dr Condescending" and "Nurse Flaky": The representation of medical practitioners in an infertility corpus

Karen Donnelly
Lancaster University
k.donnelly@lancaster.ac.uk

1 Introduction

Despite the multiplicity of studies on infertility in social science (Greil et al, 2010), there is currently little linguistic research into this topic, particularly in the UK. The majority of medical sociologists who have carried out studies in this field would agree that there are problematic implications of infertility which would merit a (critical) discourse based approach, such as, identity and representation (Thompson, 2007; Letherby, 2002) yet the linguistic manifestations of these have not yet been closely scrutinised.

Using corpus linguistic methodology, in this study I examine the linguistic representation of medical practitioners from the perspective of those blogging about the experience of infertility and triangulate this data using texts from UK news articles and clinical websites.

This approach allows a unique insight into the natural language use of women experiencing infertility, engaging with medical treatment and those who provide it.

2 Analytical framework and method

The data for this study comprises three specially built corpora of texts on infertility including; UK newspaper articles from 2006 – 2012 containing the term infertility/infertile (5, 259, 717 tokens), websites for fertility clinics from 2012 (1, 277, 736 tokens) and UK blogs written by people experiencing infertility from 2006 – 2012 (1, 604, 725 tokens). These 3 corpora provide triangulation across text types and a perspective on infertility from the media, medical and personal viewpoints.

Initial analysis was carried out using Wordsmith Tools (Scott, 2012) to elicit the top 100 lexical keywords from each corpus, which were then grouped thematically in order to allow comparison across the 3 corpora and guide selection for further study using collocations and concordance lines.

Following Baker (2006), a corpus-assisted, discourse analytical framework was applied to this data examining keywords (significantly frequent terms), collocations (words which frequently co-occur) and concordance lines (words in context) with a particular focus on identifying linguistic traces of discourses (Sunderland, 2004), in this case

of discourses around medical practitioners. The search terms produced from the initial keyword analysis and used for the concordance study include *Doctor, Dr,* and *Nurse* which are key in all 3 corpora.

Concordance lines were used to study these keywords in context and several linguistic traces were identified pointing to a range of 'named' discourses around medical practitioners, this closer analysis also uncovered the differing linguistic manifestations of particular discourses across genres. Where concordances contained traces of multiple discourses they were coded as primary and secondary in order to make as complete a study as possible.

As the key focus of the study is to examine the lived experience of infertility through the texts written by bloggers on this topic, most attention is paid to the discourses found in this corpus. However this data was triangulated through comparison to the news and clinical corpora.

3 Some findings

Several discourses emerged from my analysis of the blogs, with of the most consistently reproduced detailed below. A key discourse to emerge was "Practitioners as caricatures", which was particularly noticeable in the naming strategies of the bloggers who frequently referred to the practitioners through nicknames relating to their characteristics such as *Dr Candour, Dr Old-School, Dr Condescending, Nurse Flaky* and *Nurse Capable*. This may be used as an anonymisation strategy but is also an example of the use of humour in response to a stressful life event.

The sometimes strained relationship between practitioners and patients in infertility clinics (Becker and Nachtigall, 1991) is realised in the discourse of "Practitioners as gate keepers" in which both doctors and nurses are portrayed as both conduits and barriers to both information and treatment regimes. Included in the concordances coded for this discourse are instances of miscommunication, frustration with the system and waiting times.

Another facet of the problematic practitioner/patient relationship is found in the discourse of "the expert patient", a phenomenon which has grown substantially in the last decade and is seen by many as a double edged sword (Fox and Ward, 2006). This discourse includes the contesting of a more traditional "Dr knows best" discourse, exemplified in both of the other corpora. It also includes examples of online health help-seeking, including requests for information from fellow bloggers.

4 Conclusion

The close analysis of concordance lines was an ideal methodology for eliciting fine grained discourse traces from the data and providing a patient perspective on practitioners. Comparison with the news and clinical corpora suggests that the blogger are engaging with contesting discourses rather than the hegemonic discourses drawn on in media and medical texts. The blog texts were more likely to include negative representations of practitioners, different expectations of doctors than nurses (often presented in ways which reprised traditional gender roles) and communication between patients and practitioners as problematic and unsatisfying from a patient perspective.

References

Baker, P. 2006. *Using Corpora in Discourse Analysis.* London: Continuum

Becker, G and Nachtigall, R.D. 1991. Ambiguous responsibility in the doctor-patient relationship: The case of infertility, In: *Social Science & Medicine,* 32: 8, pp 875-885,

Fox, NJ. 2006. Health Identities: From Expert Patient to Resisting Consumer. In: *Health*, 10 (4): 461-479

Greil, A, Blevins-Slauson, K and McQuillan, J. 2010. The experience of infertility: A review of recent literature. In: *Sociology of Health & Illness* 32:1 pp. 140–162

Letherby, G. 2002. Challenging Dominant Discourses: identity and change and the experience of 'infertility' and 'involuntary childlessness. In: *Journal of Gender Studies*, 11:3 pp. 277-288

Sunderland, J. 2004. *Gendered discourses*. Basingstoke: Palgrave Macmillan.

Thompson, C. 2007. *Making parents: the ontological choreography of reproductive technologies*. Cambridge, Mass. ; London : MIT

Class matters: press representations and class distinctions in British broadsheets

Alison Duguid
University of Siena
alison.duguid@unisi.it

1 Introduction

Class is a key concept in social studies. It is a demographic and socioeconomic description used by sociologists and economists to describe groups of people. The 2013 State of the Nation Report to Parliament in the UK mentions eight socio-economic class divisions, and accepts the idea of class as defined by parental occupation; while the words *working class* are used the term *middle class* is never mentioned. The report finds that class is a bigger barrier than gender to getting a top job; that the class effect is bigger than the gender effect. It also declares that:

> Opportunities to move between different social classes or to a higher income group – absolute social mobility – matters as an indicator of society's capacity to enable people to get on in life.
> As birth not worth has become more a determinant of life chances, higher social mobility – reducing the extent to which a person's class or income is dependent on the class or income of their parents – has become the new holy grail of public policy.
> *(Hills et al 2010)*

In essence the report deals with the question of social mobility or, rather, the lack of it. Class is obviously an important factor in questions of social policy. Wage disparity, educational disparity, postcode disparity, social mobility, have all been in the headlines over the past year. In particular the topic of social mobility has been under discussion with much talk of the effects of inequality on the wellbeing of society as a whole. However, it is not so much the facts of the existence of diversity, disparity and inequality that are of interest to the investigating linguist but rather the way certain diversities are construed and constructed discursively. This study looks at the way in which class is handled in two British broadsheets offering a series of snapshots represented by a time defined corpus, revealing some of the pervasive meanings that construct identity.

2 Previous studies

Questions of social groupings, diversity and discrimination have been investigated many times in corpus studies and a number of analyses of the representations of minority groups have employed techniques from corpus linguistics. Krishnamurthy (1996) investigated media discourses around the words racial, ethnic, and tribal; Baker (2004) used the debates over a Bill in the House of Lords to equalize the age of sexual consent for gay men with that for heterosexuals; Duguid (2010) investigating the most frequent words prefixed by anti in 1993 and in 2005, found that the items anti-semitism and anti-semitic retained exactly the same place in the rankings and concluded that discourses in the newspapers relating to anti-semitism had remained frequent and statistically consistent. Partington (2012) followed up on this work and analysed the discourses relating to anti-semitism in three leading British national "quality" newspapers from 1993 to 2009, showing the way anti-semitism is represented, by reporting and by discussion in the UK broadsheets, and how these representations have changed over time. Baker (2010) investigated the representation of Islam in broadsheet and tabloid newspapers in the UK, Marchi (2010) carried out an MD-CADS analysis of changes in the way the British press treated issues of 'morality', including attitudes towards gay people, between 1993 to 2005; Baker and McEnery (2005), Baker et al (2008), Gabrielatos and Baker (2008), all provided analyses of the portrayal of refugees, asylum seekers and (im)migrants (collectively 'RASIM') in the British press. Morley and Taylor (2012), concerned with representations of non-EU immigrants in the British and the Italian press, found that the negative representation of immigrants occurs in the Italian corpus but not in the UK data. On the topic of ageing, Mautner (2007), searched large computerized corpora for lexico-grammatical evidence of stereotypical constructions of age and ageing. Duguid (2014) also examined the ways in which age was represented in a British newspaper corpus. But class as the object of study is rarely treated.

Class does feature in corpus studies. In the IJCL corpus of articles, reviews and editorials in the *International Journal of Corpus Linguistics* 2000-2011[29] of 1284 occurrences of the item *class* there are 36 occurrences of *social class* and 28 of *socio economic class*). The question of class is touched on: in studies for example of the British National Corpus, where four classes are distinguished, but these are mostly in terms of particular language features being preferred by one class or another (e.g. Rayson et al 2000, Berglund 2000, Deutschmann 2006, Xiao and Tao 2007); even a book about the

[29] Compiled at Bologna University by the SiBol group.

use of corpora for sociolinguistic study (Baker 2010) contains little reference to class.

3 Methodology

This preliminary study looks at some of the ways in which class itself is represented by means of search-word initiated investigation of a newspaper corpus. As part of an ongoing CADS project which aimed to investigate the way class was dealt with in the broadsheets, this study looks at some of the terms which explicitly refer to class to see how the representation has changed over the past 20 years. It concentrates on the patterns of evaluation by examining collocational profiles and textual preferences of a few chosen phraseologies using the SiBol corpus. Corpus studies allow us to identify widespread patterns but also infrequent though interesting examples, both of which may be overlooked in a small-scale analysis. The approach, (favoured by Stubbs 1996), starts with an intuitively compiled list of words dealing with social class, and then uses corpus tools to understand more about their meanings. Collocates are used to sort the data into homogeneous groups and make any recurrent patterns visible.

General newspaper corpora can provide a lens for viewing changing attitudes. The SiBol/Port corpus tracks three British broadsheets over 20 years.[30] The corpus consists of the Guardian, The Times, the Telegraph and the Sunday Times and the Sunday Telegraph from 1993; a sister corpus, containing the complete set of articles in the same newspapers (plus the Guardian's sister paper, the Observer) from 2005; a third corpus compiled by Taylor, contained the output of the Guardian, Times, Telegraph for 2010. A further collection for 2013 is being compiled extending the range of newspapers. I have used here the SiBol and Port corpora for the years 1993, 2005, 2010 and preliminary, yet to be cleaned up, versions for 2013 and 2014 interrogating the Guardian and Telegraph partitions of the corpus across time. The two papers were chosen as representing the liberal and conservative quality papers and the corpus was given the name *G and T* (see Table 1).

G and T corpus (Guardian and Telegraph	
Year	**Tokens used for word list**
1993	59,406,020
2005	87,461696
2010	89,030,576
2013	67,062,160

Table 1 G and T corpus.

[30] For SiBol see http://www3.lingue.unibo.it/blog/clb

Class and *classless* are key words for *G and T 1993* in comparison with the other years. The preliminary search revealed salient contextual elements, taken up in news discourse or opinion pieces, pinpointing periods of increased reporting on the topic. Among the search words used in the investigation were: *social mobility, class, classes, classless, middle-class, upper-class, working-class* and *posh.* Like the vast majority of corpus-based/assisted studies we use keyness analysis: frequency comparisons and collocation analysis, collocates being grouped into semantic sets.

4 Initial conclusions

The search revealed systematic patterns of presentation across a large number of texts, but also across time since the corpus is also a diachronic one. Although sociologists and economists no longer use the three-term distinction, the broadsheets still maintain them as the principle class identities. Discussion of social mobility is on the increase but the use of class terms remains steady or is decreasing. The economic situation of the country and austerity policies are represented in terms of class, in particular with reference to benefits, education and housing. Speech is still represented as a great signifier of class, reference to accents and voices are frequent in talk about class. The broadsheets are considered to be middle class media; they are often criticised by politicians and political commentators, for class preference in their hiring practices, for representing the Westminster village, for wearing north London intellectual blinkers, for not understanding the country anymore but also for being a middle class institution. Our data certainly backs up the perceived London and Westminster preoccupations in the discussion of class and it shows that discussion of the working class often occurs with reference to the regions. It also indicates that the two broadsheets under investigation tend to be judgmental rather than descriptive of the middle class, which is the butt of humour, disparaging comments, jokes and mocking references, while they are generally descriptive of the working class, which, when evaluated, tends to be evaluated positively. The tone of the broadsheets presupposes a reader who recognises their allusions, references, citations. If we take the broadsheets to be essentially middle class and the same for their readers, it would seem that self-deprecation is the main strategy in the representation of class.

References

Baker, P. 2010. Sociolinguistics and Corpus Linguistics. Edinburgh: Edinburgh University Press.

Baker, P. 2004. “Unnatural acts': Discourses of homosexuality within the House of Lords debates on gay male law reform”. Journal of Sociolinguistics 8 (1): 88-106.

Baker, P. 2010. “Representations of Islam in British broadsheet and tabloid newspapers 1999-2005”. Journal of Language and Politics. 9 (2): 310-338.

Baker, P. and McEnery, T. 2005. “A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts”. In Journal of Language and Politics 4 (2): 197-226.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., McEnery, A. M. & Wodak, R. 2008. “A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press”. Discourse and Society 19 (3): 273-306.Berglund, Ylva. 2000. “Gonna and going to in the spoken component of the British National Corpus”. *Language and Computers* 33: 35-50.

Buxton, J., Clarke, L., Grundy, E., & Marshall, C. E. 2004. “The long shadow of childhood: associations between parental social class and own social class, educational attainment and timing of first birth; results from the ONS Longitudinal Study”. *Population trends*, 121: 17-26.

Deutschmann, Mats. 2006. "Social variation in the use of apology formulae in the British National Corpus." *Language and Computers* 55 (1): 205-221.

Duguid, A. 2010. “Investigating anti and some reflections on Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS)”. *Corpora* 5 (2): 191–220.

Duguid. 2014. Old and Young: changing evaluations of intergenerational diversity. In G. Balirano, M.C. Nisco (eds), *Languaging Diversity: Identities, Genres, Discourses*. Newcastle Cambridge Scholars Publishing.

Gabrielatos, C. and Baker, P. 2008. “Fleeing, Sneaking, Flooding: A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the UK Press, 1996-2005”. *Journal of English Linguistics*, *36*(1): 5-38.

Hills, J., Brewer, M., Jenkins, S. P., Lister, R., Lupton, R., Machin, S. & Riddell, S. 2010. *An anatomy of economic inequality in the UK:* Report of the National Equality Panel.

Krishnamurthy, R. 1996. “*ethnic*, *racial*, and *tribal*. The language of racism?” In C. Caldas-Coulthard and M. Coulthard (eds.) *Texts and Practic.es: readings in Critical Discourse Analysis.* London and New York: Routledge.

Marchi, A. 2010. “The moral in the story: a diachronic investigation of lexicalised morality in the UK press”. *Corpora* 5 (2): 161-189.

Mautner, G. 2007. “Mining large corpora for social information: The case of elderly”. *Language in Society* 36 (01): 51-72.

Morley, J. and Taylor, C. 2012. “Us and them: how immigrants are constructed in British and Italian newspapers”. In P.Bayley and G. Williams (eds.*) European Identity what the media say*. Oxford: OUP

Partington, A. 2012. “The changing discourses on anti-semitism in the UK press from 1993 to 2009: A modern-diachronic corpus-assisted discourse study”. *Journal of Language & Politics*11 (1): 51-76.

Savage, M., Devine, F., Cunningham, N., Taylor, M., Li, Y., Hjellbrekke, J., and Miles, A. 2013. “A new model of social class? Findings from the BBC’s Great British Class Survey Experiment.” *Sociology*, *47*(2): 219-250.

Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.

Taylor, C. 2013. ‘Searching for similarity using corpus-assisted discourse studies.’ *Corpora* 8 (1): 81-113.

Xiao, R. and Tao, H. 2007. “A corpus-based sociolinguistic study of amplifiers in British English”. *Sociolinguistic Studies* 1(2): 241-273.

Designing and implementing a multilayer annotation system for (dis)fluency features in learner and native corpora

Amandine Dumont
Université catholique de Louvain
amandine.dumont@uclouvain.be

1 Introduction

The notions of fluency and disfluency have mainly been approached from two angles. The holistic approach defines fluency as the "smooth, rapid, effortless use of language" (Crystal 1987: 421); in this perspective, fluency is considered as a global phenomenon of language production (e.g. Chambers 1997; Lennon 1990). By contrast, the componential approach (Fox Tree 1995; Gut 2009; Hieke 1981, 1985) sees fluency as an isolatable dimension of language proficiency resulting from the conjunction of a series of quantifiable and qualifiable phenomena, such as filled pauses, discourse markers, false starts or restarts. Many componential studies are, however, limited to the research of one of those features without considering its interaction with other fluency features (Chafe 1980; Pallaud et al. 2013; Raupach et al. 1980), and few have considered variation in speech performances. The researcher's choice for one or the other of those two approaches obviously results in a different set of fluency features and in diverging methodologies.

Against this backdrop, Götz' (2013a) study is remarkable for combining both perspectives into what she calls an "integrated" approach: she examined a comprehensive set of fluency variables with the aim of delineating the overall fluency profiles of nonnative speakers of German. Following this new line of thinking, this paper aims to present an innovative multilayer annotation system for (dis)fluency features in learner and native speech. This system allows for the investigation of a large number of (dis)fluency features (either independently or in interaction) with a view to drawing the bigger picture of (dis)fluency behavior of nonnative and native speakers. The system has been created within a wider research project on fluency across languages and modalities[31] for which a more general framework of (dis)fluency annotation has been developed (Crible et al. 2014).

2 Design

Several theoretical and practical principles have underpinned the design of the (dis)fluency annotation system.

The main hypothesis of the research project out of which this annotation system has arisen is that fluency and disfluency are the two sides of the same coin. In other words, the same feature can be used as a means to enhance fluency at one point, and as a marker of disfluency at another, and it is in the recurrence and combination of those features that fluency or disfluency can be established. Consequently, the tagging system makes no a priori decision as to which elements should be considered as fluent or disfluent: all occurrences of a feature are tagged in the same way.

The integrated approach to (dis)fluency, i.e. (dis)fluency seen as a variety of features contributing to a holistic phenomenon, constitutes the second cornerstone of the system. For this purpose, the protocol offers a tagging system for a dozen distinct (dis)fluency features (see Table 1). It allows for the annotation of (dis)fluency features involving one (e.g. a pause) or several words (typically a repetition), and, conversely, for the annotation of words included in more than one (dis)fluency feature (e.g. a vowel lengthening within a repetition). The annotation system remains essentially linear, and (dis)fluency features are annotated at the level of the word. In addition to this componential dimension, the system makes it possible to draw a holistic picture of each individual speaker's fluency behavior.

Thirdly, the system is designed for and on the basis of spoken data: contrarily to some other annotation systems which have been developed with standard written text in mind (see Rehbein et al. 2012), this protocol is solely based on concepts of spoken language such as filled pause, self-correction, aborted utterance and the use of "written" concepts such as "sentence" is avoided. Preliminary versions of the annotation scheme were iteratively tested on a corpus sample and amended accordingly to reach the final version.

Last but not least, the (dis)fluency annotation system is aimed to be applied to large corpora, to different speaking tasks, and to both learner and native data. This implies that the system must not only be grounded on well-defined (dis)fluency categories, but it also has to be flexible, straightforward, applicable to different data types and reasonably quick to implement.

3 Implementation

This multilayered (dis)fluency annotation system has been implemented within the EXMARaLDA tool (Schmidt et al. 2009) to the time-aligned version of the French component of the *Louvain International Database of Spoken English Interlanguage*

[31] Université catholique de Louvain & Université de Namur (ARC grant 12/17-044).

(LINDSEI, Gilquin et al. 2010).

LINDSEI (inline annotations)	(DIS)FLUENCY ANNOTATION SYSTEM	
	(Dis)fluency feature	Examples (FR009, FR010 & FR011[32])
Empty pause (perceptive transcription)	Unfilled pause (in ms; 3 sub-categories)	*I'd been* ***(0.720)*** *planning to to go*
Filled pause	Filled pause	*something to do with* ***er*** *politics*
Truncated word	Truncated word (3 sub-categories)	***wh when*** *I was a little girl*
Foreign word	Foreign word	*politics or (0.820)* ***relations internationales***
Lengthening	Lengthening	*in* ***a*** *(0.280) well in a real (0.750) town*
/	False start	***I don't think it's*** *well for me er I wouldn't do it*
/	Repetition	*it's just* ***for*** *(0.490)* ***for*** *me* ***it's it's*** *meant for students I think*
/	Restart (5 sub-categories)	*last week (0.190) er* ***last year*** *he plays with the (0.530)* ***he plays volleyball with the Lux***
/	Connector	*I like skiing too* ***but*** *it's a bit too far to go* *you have erm (0.570) facilities* ***and*** *you can get into contact with people*
/	Discourse marker	***well*** *I (0.200) wanted to to go there* *I like eh making er swimming for (0.360) non stop* ***you know***
/	Editing term	*tennis table (0.330) ta table tennis* ***sorry***

Table 1. LINDSEI mark-up vs. (dis)fluency annotation system

This large database contains recordings and transcripts of interviews of advanced learners of English from 11 mother tongue backgrounds (50 interviews [c. 20 min. each] per L1, each interview consisting of three speaking tasks)[33]. Although the released version of LINDSEI transcriptions contains inline annotations of several features of spoken language (including (dis)fluency phenomena), these are insufficient for (dis)fluency analyses answering the principles outlined above. Table 1 illustrates the added value such annotation system can provide to spoken corpora.

Each feature has a corresponding tag in the form of one or two letters, e.g. FP for filled pause and T for truncated word. For repetitions, a numbering system is used to show the number of repetitions and the number of repeated words. A set of symbols is also integrated to indicate the onset ("<") and offset (">") of each feature as well as multiple tagging on one item, if any ("+"; e.g. the word "enfin" in *it was erm enfin we hadn't* [FR005], which is tagged both as a discourse marker and as a foreign word). Those tags are spread into successive layers of annotation, corresponding to different levels of precision in the characterization of (dis)fluency features, from the more generic to the more in-depth. The following examples[34] illustrate the annotation system.

- FR017

er	(1.590)	you	we	have	to
<FP>	<UP>		<RS>		<L>
	<P>		<SP>		

study	erm	mathematics	er	to	start
	<FP>		<FP>		

- FR027

I	(0.220)	well	(0.690)	I	won	't
<R0	<UP>	<DM>	<UP>	R1>		
	<N>+<S>	<N>	<N>+<P>			

say	I	(0.120)	only	listen	to	classical	music
		<UP>					
		<S>					

- FR011

a	girl	sitting	in	a	(2.030)	yes
			<R0	R0	<UP>	<DM>
					<N>+<P>	<N>

[32] Each interview in LINDSEI is identified by a specific code: "FR" corresponds to the interviewees' mother-tongue (here French), and the three-figure number (001 to 050) refers to the 50 learners.

[33] A comparable corpus of interviews of native speakers of English, the *Louvain Corpus of Native English Conversation* (LOCNEC, De Cock 2004) is currently being time-aligned and annotated for (dis)fluency features so as to provide a proper native benchmark.

[34] 1st tier: FP: filled pause; UP: unfilled pause; RS: restart; L: lengthening; R*n*: repetition; DM: discourse marker; T: truncated word. 2nd tier: S: short UP; P: long UP; SP: propositional substitution; N: nesting.

in	a	(0.230)	o	on	a	chair
R1	R1>	<UP>	<RS+<T	RS+T>	RS>	
		<S>	<SP	SP>		

In order to highlight the potential of this type of annotation, I adopt a contrastive approach to test the hypothesis that learners and native speakers differ in their quantitative use of both (dis)fluency features and (dis)fluency patterns/clusters (cf. Aijmer 1997). On the basis of the annotated LINDSEI-FR and its native speaker counterpart LOCNEC, and of the features outlined above, (dis)fluency profiles of 30 learners as compared to 30 native speakers are presented in an integrated approach to (dis)fluency. Preliminary data reveal that advanced French learners of English are a much less homogeneous group than native speakers as the extent of inter-speaker variation is far greater (about twice as large), particularly in the use of filled and unfilled pauses. The paper then shifts to the qualitative use of those two types of pauses and examines how native and nonnative speakers compare with respect to (dis)fluency clusters around those pauses. This study offers an L1/L2 counterpart to Degand & Gilquin's (2013) recent research on the environment of pauses in L1 English and French and contributes to the recent line of studies into variation between learners (from the same or different mother tongue, or with different language proficiency levels) on the one hand, and between native and nonnative speakers on the other (e.g. Gilquin & Granger 2011; Gilquin & De Cock 2011; Götz 2013 a and b).

References

Aijmer, Karin. 1997. "'I Think' - an English Modal Particle". In *Modality in Germanic Languages. Historical and Comparative Perspectives*, eds. Toril Swan & Olaf J. Westvik, 1-47. Berlin: Mouton de Gruyter.

Chafe, Wallace. 1980. "Some Reasons for Hesitating". In *Temporal Variables in Speech*, eds Raupach, Manfred et al, 168-80. Den Haag: Mouton de Gruyter.

Chambers, Francine. 1997. "What Do We Mean by Fluency?". *System* 25, n° 4: 535-44.

Crible, Ludivine, Dumont, Amandine, Grosman, Iulia, & Notarrigo Ingrid. (2014). "Annotation des marqueurs de fluence et disfluence dans des corpus multilingues et multimodaux, natifs et non natifs". Unpublished internal report. Université catholique de Louvain: Louvain-la-Neuve.

Crystal, David. 1987. *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.

De Cock, Sylvie. 2004. Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL)*, New Series 2, 225–246.

Degand, Liesbeth, & Gaëtanelle Gilquin. 2013. "The clustering of 'fluencemes' in French and English". 7th International Contrastive Linguistics Conference (ICLC 7) - 3rd conference on Using Corpora in Contrastive and Translation Studies (UCCTS 3) (Ghent, 11/07/2013 - 13/07/2013).

Fox Tree, Jean E. 1995. "The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech". *Journal of Memory and Language* 34: 709-38.

Gilquin, Gaëtanelle, & Sylvie De Cock. 2011. "Errors and Disfluencies in Spoken Corpora: Setting the Scene". *International Journal of Corpus Linguistics* 16, n° 2: 141-72.

Gilquin, Gaëtanelle, Sylvie De Cock, & Sylviane Granger, eds. 2010. *LINDSEI. Louvain International Database of Spoken English Interlanguage*. Presses Universitaires de Louvain. Louvain-la-Neuve.

Gilquin, Gaëtanelle, & Sylviane Granger. 2011. "The Use of Discourse Markers in Corpora of Native and Learner Speech: From Aggregate to Individual Data". Corpus Linguistics conference (Birmingham, 20/07/2011 - 22/07/2011).

Götz, Sandra. 2013a. *Fluency in Native and Nonnative English Speech*. Studies in Corpus Linguistics (SCL) 53. Amsterdam & Philadelphia: John Benjamins Publishing Company.

——. 2013b. "How Fluent Are Advanced German Learners of English (perceived to be)? Corpus Findings vs. Native-Speaker Perception". In *Studies in Variation, Contacts and Change in English*, eds. Magnus Huber & Joybrato Mukherjee, Vol. 13. Giessen: University of Giessen.

Gut, Ulrike. 2009. *Non-Native Speech: A Corpus-Based Analysis of Phonological and Phonetic Properties of L2 English and German*, eds. Thomas Kohnen & Joybrato Mukherjee. English Corpus Linguistics 9. Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien: Peter Lang.

Hieke, Adolf E. 1981. "Audio-Lectal Practice and Fluency Acquisition". *Foreign Language Annals* 14, n° 3: 189-94.

Hieke, Adolf E. 1985. "A Componential Approach to Oral Fluency Evaluation". *The Modern Language Journal* 69 (2): 135-42.

Lennon, Paul. (1990). "Investigating Fluency in EFL: A Quantitative Approach*". *Language Learning* 40, n° 3: 387-417.

Pallaud, Bertille, Stéphane Rauzy, & Philippe Blache. 2013. "Auto-interruptions et disfluences en français parlé dans quatre corpus du CID". *TIPA: Travaux interdisciplinaires sur la parole et le langage*, n° 29.

Raupach, Manfred, Hans-Wilhelm Dechert, & Frieda Goldman-Eisler, eds. 1980. *Temporal Variables in Speech*. Janua Linguarum 86. The Hague: Mouton.

Rehbein, Ines, Sören Schalowski, & Heike Wiese. 2012.

"Annotating Spoken Language". In *Best Practices for Speech Corpora in Linguistic Research*, 29. Istanbul, Turkey.

Schmidt, Thomas, & Kai Wörner. 2009. "EXMARaLDA - Creating, Analyzing and Sharing Spoken Language Corpora for Pragmatics Research". *Pragmatics - Quarterly Publication of the International Pragmatics Association* 19 (4): 565. www.exmaralda.org.

Traitor, whistleblower or hero? Moral evaluations of the Snowden-affair in the blogosphere

Dag Elgesem
University of Bergen
dag.elgesem@uib.no

Andrew Salway
Uni Research
andrew.salway@uni.no

1 Introduction

The mining of social media data both for commercial purposes by private actors and for surveillance purposes by government agencies is becoming increasingly widespread and invasive (van Dijck, 2014). Since people now spend much of their social life on social media, whilst having limited control over how their data are used, the legitimacy of these practices is a critical social issue. To better understand the social consequences of this extensive mining of personal data it is important to investigate how social media users themselves perceive of social media "dataveillance". Previous research into what users think of social media mining showed that while most people are concerned about how information about themselves is used, there are significant differences in knowledge among users, as well as differences in the kinds of concern they have (Kennedy et.al. 2015).

In 2013 Edward Snowden made public information about a number of surveillance programs, including the so-called PRISM program, a NSA program to systematically tap into user data from social media platforms like Facebook, Twitter, Skype, and LinkedIn (Greenwald, 2014). Public opinion is divided over the political and moral issues raised by these leaks. In a survey undertaken by the PEW-center, 45% of the respondents agreed that "Snowden's leaks of classified information about programs has served public interest" while 45% agreed that the leaks have "harmed public interest" (PEW, 2014). Interestingly, the same survey showed that 55% agreed that the government should pursue a criminal case against him while 31% disagreed with this (PEW, 2014). This suggests that some people think both that he did a service to society and that he should be brought to justice.

The differences in opinions over Snowden and his acts were also visible in the editorial policies of newspapers (Greenwald, 2014). Newspaper editors discussed how to frame Snowden in their coverage of the affair; some newspapers called him a "whistleblower" – perhaps implying that he was a person who has exposed wrongdoing, while others referred to him more negatively as a "leaker" or just a "source" (Wemple, 2013).

The Snowden affair triggered extensive debates about the legitimacy of the PRISM program and the other surveillance activities he leaked information about. The affair also gave rise to discussions about the moral evaluation of what Snowden did; i.e. his disclosure of graded information and subsequent flight to Russia.

In this paper we take a corpus-based discourse analysis approach in order to investigate how bloggers discussed and evaluated surveillance and the PRISM program, with a focus on how the two aspects mentioned above – legitimacy of surveillance/PRISM and moral evaluation of Snowden's actions – are related in the bloggers' discourse. It seems plausible to expect that people who use information revealed by Snowden as a basis for criticism of the surveillance programs will have a tendency to also evaluate Snowden's actions positively. But if this is the case, will they also express their moral evaluations of Snowden in the course of an argument about the PRISM program? Perhaps, since Snowden as a person is so controversial, people who want to critically discuss the PRISM program will try to avoid taking a stance on the moral status of his acts, in order not to detract attention from the main issue?

Thus our research question is: How do bloggers discuss and evaluate the legitimacy of the PRISM program and, in particular, does their discourse about the PRISM program also involve a moral evaluation of Snowden?

2 Method

A corpus of approximately 100,000 English-language blog posts related to the topic of surveillance was gathered by daily querying of three search engine APIs (Google, Bing and Yahoo). Twenty-one query terms were chosen based on domain expertise and inspection of frequent n-grams in some relevant blog posts, e.g. "data retention", "edward snowden", "electronic surveillance", "fisa court", "government surveillance", "intelligence agencies", etc. The queries were restricted to three blog platforms – WordPress, Blogspot and Typepad – which analysis had predicted would include the vast majority of relevant posts. The collected posts were processed with JusText[35] to extract the main text content, which was stored along with the date (month and year) that was extracted from the URL. From this corpus, blog posts containing "Snowden", and dated from June 2013 (when *The Guardian* first published the classified information) to June 2014, were selected for analysis, i.e. approximately 15,000 posts.

A ranked list of collocates for "Snowden" was generated (span 5 words, ranked by pointwise mutual information) in order to identify words that seem to be commonly used in moral evaluations. Blog posts containing these words were then subject to concordance, collocation and word cluster analyses in order to investigate how Snowden and his actions were morally evaluated, i.e. "whistleblower" (2,683 blog posts), "leaker" (1,026), "traitor" (928) and "hero" (887).

Secondly, blog posts containing the words "surveillance" (8,213 blog posts) and "PRISM" (2,113) were analysed manually to investigate how the credibility of such activities and programs was evaluated. Random samples of 100 blog posts for "surveillance" and "PRISM" were coded and analysed. The coding scheme distinguished between posts where the blogger expressed a critical opinion of the PRISM program ('subjectively critical'), and blog posts that only wrote about problematic aspects of the program but did not express a personal criticism ('objectively critical'). A similar distinction was made on the positive side ('subjectively supportive' and 'objectively supportive'). Some further posts were coded as 'neutral'. Additionally we recorded whether the posts expressed a negative evaluation of Snowden, a positive moral evaluation, or were neutral in their reference to Snowden.

3 Main findings

We find that the term "traitor" is used mostly to express a negative moral evaluation of Snowden. Word cluster analysis around the 1,468 instances of "traitor" shows the most frequent constructions are used to depict him as a traitor. There are however also instances where the term is used in discussions about whether Snowden should be characterized as "traitor or hero?". A closer inspection of a sample of posts that discuss this question showed that many of them either do not take a stance, or are positive to him. The same is true of blog posts containing the term "hero". Again, "hero" is in most cases used to express a strongly positive evaluation but, in some cases, it is used in the course of a discussion of what the right moral evaluation of Snowden is. The term "leaker" is however mostly used in contexts where Snowden is described in otherwise neutral terms. Perhaps surprisingly, an occurrence of the term "whistleblower" is not a clear signal of a positive moral evaluation of Snowden since a majority of bloggers seem to use the term in a more technical sense.

Regarding how surveillance activities and programs are discussed in the blogs we found that the majority of the coded blog posts were either subjectively or objectively critical of the PRISM program and surveillance. Interestingly, the vast

[35] https://code.google.com/p/justext/

majority of these blog posts mentioned Snowden in a neutral way and did not make a clear moral evaluation of him, i.e. there were rather few blog posts which both morally evaluated Snowden and evaluated the credibility of surveillance and PRISM.

4 Discussion and conclusions

The first part of our research question is how bloggers evaluate the legitimacy of the PRISM program. Here we find that most bloggers either express their personal disapproval of the surveillance program, or report on PRISM from a critical perspective. This is consistent with previous research on users' attitudes to social media mining: when considering concrete measures of surveillance they tend to express concern (Kennedy et. al. 2015).

Our analyses suggest, however, that the critical discussions of the PRISM program and the policies of which it is a part, are in most cases not combined with a positive moral evaluation of Snowden. The bloggers typically separate the two issues by referring to Snowden in neutral terms. Furthermore, few of the blogs that dismiss Snowden as a "traitor" or endorse him as a "hero" discuss the PRISM program. Our answer to the second part of our research question is thus that in most cases the bloggers' discourse on PRISM do not also involve a moral evaluation of Snowden.

In ongoing work we further explore the structures of the bloggers' discourse on the Snowden affair by also considering the blogs' networks of links to other blogs and websites. This contextual information can provide more insight into the discursive situation the individual blog posts are written in response to and thus give us a better understanding of the bloggers' perception of both social media mining and the Snowden affair.

Acknowledgements

This research was supported by a grant from the Research Council of Norway's VERDIKT program (NTAP, project 213401). We are very grateful to Knut Hofland for his role in creating the corpus analysed here.

References

Greenwald, G. 2014. *No Place to Hide. Edward Snowden, the NSA and the Surveillance State*. London: Hamish Hamilton.

Kennedy, H., Elgesem, D. and Miguel, C. 2015. "On fairness: User perspectives on social media mining". Forthcoming in Convergence

PEW Research Center. 2014. "Most young Americans say Snowden has served the public interest". http://www.pewresearch.org/fact-tank/2014/01/22/most-young-americans-say-snowden-has-served-the-public-interest/

van Dijck, J. 2014. "Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology". Surveillance & Society 12(2):197-208.

Wemple. E. 2013. "Leaker, Source or Whistleblower". http://www.washingtonpost.com/blogs/erik-wemple/wp/2013/06/10/edward-snowden-leaker-source-or-whistleblower/

Corpus statistics: key issues and controversies (*panel*)

Stefan Evert
FAU Erlangen-Nürnberg
stefan.evert@fau.de

Gerold Schneider
University of Zürich
Gschneid@es.uzh.ch

Vaclav Brezina
Lancaster University
v.brezina@lancaster.ac.uk

Stefan Th. Gries
University of California, Santa Barbara
stgries@linguistics.ucsb.edu

Jefrey Lijffijt
University of Bristol
jefrey.lijffijt@bristol.ac.uk

Paul Rayson
Lancaster University
p.rayson@lancaster.ac.uk

Sean Wallis
University College London
s.wallis@ucl.ac.uk

Andrew Hardie
Lancaster University
a.hardie@lancaster.ac.uk

1 Motivation

The application of sound statistical techniques in descriptive linguistics is increasingly seen as a vital methodological requirement. However, there are still many studies that fail to carry out a statistical analysis or, more commonly, apply significance tests and other well-established methods in an overly simplistic manner. Typical examples are significance testing of frequency differences with a chi-squared or Fisher exact test instead of multifactorial models; the exclusive use of p-values, disregarding effect size; and the visualization of keywords in the form of word clouds (which are particularly popular in the digital humanities community).

There are various reasons for this problem: researchers may not be aware of an appropriate statistical test, they may not have the tools to execute that test, or it may be an open scientific question which test would be most applicable. Accordingly, there is an urgent need for discussions about the appropriate use of statistics in quantitative linguistic studies, the development of new methods and appropriate software tools, and the dissemination of new methodological findings to the corpus linguistics community.

2 Speakers

The panel discussion brings together researchers who are well known for their research on statistical methodology, their teaching efforts in this area and/or the implementation of relevant software tools. Conference delegates will gain a deeper understanding of key problems and learn about the latest methodological developments.

3 Format and topics

We have defined a list of five key topics for the panel. Two panellists are invited to give position statements on the topic, sketching opposite points of view or suggesting alternative solutions. This is followed by a discussion among panellists. We then invite comments and questions from the audience.

4 Experimental design – which factors should we measure?

Recent work has shown that simple frequency comparisons and similar approaches are inappropriate in most cases (e.g. Evert 2006). Instead, multifactorial models could be used (Gries 2006) in order to account in full for the variability of frequency counts and other measures, or the data could be modelled differently (Lijffijt *et al.* 2014). Key questions to be discussed are (i) the unit of measurement and (ii) which predictive factors should be included in the analysis. Regarding the unit of measurement, should studies report and model per-word counts or per-text relative frequencies, or rather predict the outcome of a speaker decision? In the latter case, we base our investigation on an envelope of variation (Labov 1969), such as an alternation, and are potentially less affected by corpus sampling. When selecting a set of predictive factors, we need to strike a reasonable balance between too few, which runs the risk of excluding important factors and thus resulting in an unsatisfactory goodness-of-fit, and too many, which leads to sparse data problems, overadaptation of the model to the data set, and limited scientific insights.

5 Non-randomness, dispersion and violated assumptions

"Language is never, ever, ever random" (Kilgarriff 2005). In particular, words and other linguistic phenomena are not spread homogeneously across a text or corpus (Church 2000), their appearance depending on the style and topic of a text as well as previous occurrences in a discourse. As a result, the random sample assumption underlying most statistical techniques is very often violated. For example, the individual texts comprising a corpus have usually been sampled independently, but the

word tokens within each text are correlated. Therefore, when using words as a unit of measurement, the independence assumption made by frequency comparison tests and many multifactorial models is violated. We discuss the precise assumptions of different statistical techniques, under what circumstances they are violated, which violations are most harmful, and how this problem can be solved or mitigated.

6 Teaching and curricula

Corpus linguistics employs quantitative methods that rely on correct use of different statistical procedures. It therefore necessarily presupposes a certain awareness of statistical assumptions and principles. The question, however, is to what extent corpus linguists (researchers and students) should be able to perform complex statistical procedures such as mixed effects modelling using R or similar software packages. This also raises a number of other questions:

How can we improve the understanding of basic statistics among researchers and in the linguistics curricula? Should statistics courses be compulsory at BA or MA level? And perhaps even an introduction to computer programming? We also report on our personal experiences of teaching the statistics language R to students with no previous programming experience.

7 Visualisation

In statistical textbooks, initial visualisation of the data (using scatter plots, box plots, etc.) is often recommended as an important stage of data exploration before statistical tests are applied. Indeed, good visualisation can provide us with a holistic picture of the main tendencies in the data, help to discover interesting patterns, and reveal outliers and other problematic aspects of a data set. In corpus linguistics, different visualisation techniques have been used: word clouds, word trees, collocation networks, bar charts, error bars, etc. (see, e.g., Siirtola *et al.* 2011). Which of these visualisation techniques are helpful for the researcher and the reader? Does visualisation really help the reader to understand a concept and the researcher to detect interesting patterns and crucial zones, on which to focus in further investigations? Is visualisation merely a form of presentation of the data or does it play a more fundamental role in the research process?

8 Which models can we use?

There is a large range of statistical models to choose from (e.g. Schneider 2014). In topics 3.1 and 3.2 we have already talked at length about regression models, but alternative, computationally more demanding techniques are also available, such as probabilistic models from natural language processing (taggers, parsers, machine translation, text mining tools, semantic classifiers, spell-checkers) and dimensionality reduction approaches. Both the possibilities and their complexities are vast, making this discussion topic open-ended.

References and select bibliography

Brezina, V. and Meyerhoff, M. 2014. "Significant or random? A critical review of sociolinguistic generalisations based on large corpora." *International Journal of Corpus Linguistics* 19 (1): 1–28.

Church, K. 2000. "Empirical estimates of adaptation: The chance of two Noriegas is closer to p/2 than p^2." In: *Proceedings of the 17th conference on Computational linguistics*, pp. 180–186.

Evert, S. 2006. "How random is a corpus? The library metaphor." *Zeitschrift für Anglistik und Amerikanistik* 54 (2): 177–190.

Evert, S.; Schneider, G.; Lehmann, H. M. 2013. "Statistical modelling of natural language for descriptive linguistics". Paper presentation at *Corpus Linguistics 2013*, Lancaster, UK.

Gries, S. Th. 2006. "Exploring variability within and between corpora: some methodological considerations." *Corpora* 1 (2): 109–151.

Gries, S. Th. to appear. "Quantitative designs and statistical techniques." In D. Biber and R. Reppen (eds.) *The Cambridge Handbook of Corpus Linguistics*. Cambridge: Cambridge University Press.

Kilgarriff, A. 2005. "Language is never ever ever random." *Corpus Linguistics and Linguistic Theory* 1(2): 263–276.

Labov, W. 1969. "Contraction, deletion, and inherent variability of the English copula." *Language* 45(4): 715–762.

Lijffijt, J; Nevalainen, T.; Säily, T.; Papapetrou, P.; Puolamäki, K.; Mannila, H. 2014. "Significance testing of word frequencies in corpora." *Digital Scholarship in the Humanities*, online ahead of print.

Pipa, G. and Evert, S. 2010. "Statistical models of non-randomness in natural language." Presentation at *KogWis 2010*, Potsdam, Germany.

Schneider, G. 2014. *Applying Computational Linguistics and Language Models: From Descriptive Linguistics to Text Mining and Psycholinguistics*. Cumulative Habilitation, Faculty of Arts, University of Zürich.

Siirtola, H; Nevalainen, T.; Säily, T.; Räihä, K.-J. 2011. "Visualisation of text corpora: A case study of the PCEEC." In T. Nevalainen and S. M. Fitzmaurice (eds.) *How to deal with data: Problems and approaches to the investigation of the english language over time and space* (Studies in Variation, Contacts and Change in English 7). Helsinki: VARIENG.

Explaining Delta, or: How do distance measures for authorship attribution work?

Stefan Evert
FAU Erlangen-Nürnberg, Germany
stefan.evert@fau.de

Thomas Proisl
FAU Erlangen-Nürnberg, Germany
thomas.proisl@fau.de

Christof Schöch
University of Würzburg, Germany
christof.schoech@uni-wuerzburg.de

Fotis Jannidis
University of Würzburg, Germany
fotis.jannidis@uni-wuerzburg.de

Steffen Pielström
University of Würzburg, Germany
pielstroem@biozentrum.uni-wuerzburg.de

Thorsten Vitt
University of Würzburg, Germany
thorsten.vitt@uni-wuerzburg.de

1 Introduction

Authorship Attribution is a research area in quantitative text analysis concerned with attributing texts of unknown or disputed authorship to their actual author based on quantitatively measured linguistic evidence (see Juola 2006; Stamatatos 2009; Koppel et al. 2009). Authorship attribution has applications in literary studies, history, forensics and many other fields, e.g. corpus stylistics (Oakes 2009). The fundamental assumption in authorship attribution is that individuals have idiosyncratic habits of language use, leading to a stylistic similarity of texts written by the same person. Many of these stylistic habits can be measured by assessing the relative frequencies of function words or parts of speech, vocabulary richness, and many other linguistic features. Distance metrics between the resulting feature vectors indicate the overall similarity of texts to each other, and can be used for attributing a text of unknown authorship to the most similar of a (usually closed) set of candidate authors.

The aim of this paper is to present findings from a larger investigation of authorship attribution methods which centres around the following questions: (a) How and why exactly does authorship attribution based on distance measures work? (b) Why do different distance measures and normalization strategies perform differently? (c) Specifically, why do they perform differently for different languages and language families, and (d) How can such knowledge be used to improve authorship attribution methods?

First, we describe current issues in authorship attribution and contextualize our own work. Second, we report some of our earlier research into the question. Then, we present our most recent investigation, which pertains to the effects of normalization methods and distance measures in different languages, describing our aims, data and methods..

2 Current issues in authorship attribution

There are several key elements to any authorship attribution study: the nature and extent of textual material available, the richness of metadata about the texts, the number and types of linguistic features used, the strategy used to normalize the resulting feature vectors, an optional dimensionality reduction step (often by principal component analysis), the measure used to assess distances between feature vectors, and the method for classification or clustering of the texts based on feature vectors and inter-text distances. All of these aspects are currently topics of investigation and debate in the authorship attribution community (e.g. Argamon 2008; Eder and Rybicki 2013). This paper is mainly concerned with the role of standardization and normalization of feature vectors, the choice of suitable features, and the impact of different distance metrics.

The current state of the art is to consider normalization and metric as one joint step in the process of authorship attribution. One ground-breaking measure, Burrows's Delta (Burrows 2002), can in fact be understood as a combination of standardization (i.e. z-transformation) of frequency counts combined with the well-known “Manhattan” (or “city block”) metric. Many other measures proposed in the literature also amalgamate the two steps (e.g. Hoover 2004a, 2004b; Smith and Aldridge 2011). In this paper, we follow Argamon's (2008) lead and consider normalization strategy and distance measure separately from each other. This allows us to investigate the influence of each parameter on authorship attribution results as well as the interaction of these two parameters.

3 Previous work

In recent previous work, we describe an empirical investigation of the performance of 15 different text distance measures available for authorship attribution. For evaluating their performance, we compiled three collections of novels (English, French, German), each consisting of 75 complete texts of known authorship (three novels each by 25 authors), and ranging from the early nineteenth century to the first half of the twentieth century. The texts come from Project Gutenberg, the TextGrid collection and Ebooks libres et gratuits.

We compared the performance of the different

text distance measures for feature vectors of 100–5000 most frequent words (mfw) and for all three corpora. We used two quantitative measures to evaluate performance: (a) the accuracy of the clustering results relative to the gold standard if each cluster is labelled with the appropriate author; (b) a comparison of the average distance between works of the same author with the average distance between works by different authors.

As a result, we were able to demonstrate that most modifications of Burrows's original Delta suggested in the recent literature do not yield better results, even though they have better mathematical justification. Our results indicate that Eder's Delta, a measure specifically designed for highly inflected languages, does perform slightly better on French texts. The best distance measure for authorship attribution is the cosine-based Delta measure recently suggested by Smith and Aldridge (2011). Also, most text distance measures work best if between 1000 and 2000 of the most frequent words are used (Jannidis et al. 2015).

4 Current research

This work has lead us to several further questions: First, how do the effects of normalization and distance measure interact with each other? Second, why does the performance of a given combination of normalization and distance measure vary across different languages? And can this variation be explained by looking at the frequency distributions of individual, highly frequent words across texts in different languages? Finally, how can we identify the words (or features) that contribute most to the overall distance between texts? Are there linguistic or distributional explanations why these words are particularly indicative of the authorship of a text?

We approach this set of problems from two perspectives. First, we look at some mathematical properties of the authorship classification problem, based on geometric and probabilistic interpretations of the text distance measures. Argamon (2008) suggests two versions of Delta that can be interpreted in terms of statistical significance tests. However, our previous empirical results show that they are inferior to other measures that lack a similarly well-founded mathematical motivation. We are currently investigating the reasons for this discrepancy, with a particular focus on the role of different normalization strategies and their interaction with various distance measures. The results will show which aspects of the word frequency profiles of text samples are exploited by successful authorship classification methods. They may also help to identify salient lexical features that distinguish the individual writing styles of different authors.

Second, we explore another strategy for obtaining the set of features. Instead of relying on a specified number of most frequent words (mfw), we systematically identify a set of discriminant words by using the method of recursive feature elimination. We repeatedly train a support vector classifier and prune the least important features until we obtain a minimal set of features that gives optimal performance. The resulting feature set is much smaller than the number of mfw typically required by Delta measures. It contains not only function words but also common and not so common content words. The features work well on unseen data from the same and from different authors, not only yielding superior classification results, but also outperforming the mfw approach for clustering texts. This preliminary finding stands in contrast to accepted stylometric lore that function words are the most useful feature for discriminating texts from different authors.

References

Argamon, S. 2008. "Interpreting Burrows' Delta: Geometric and probabilistic foundations." *Literary and Linguistic Computing* 23(2), 131–147.

Burrows, J. 2002. "'Delta' – A measure of stylistic difference and a guide to likely authorship." *Literary and Linguistic Computing* 17(3), 267–287.

Eder, M. and Rybicki, J. 2013. "Do birds of a feather really flock together, or how to choose training samples for authorship attribution." *Literary and Linguistic Computing* 28(2), 229–236.

Juola, P. 2006. "Authorship Attribution." *Foundations and Trends in Information Retrieval* 1(3), 233–334.

Jannidis, F; Pielström, S.; Schöch, C.; Vitt, Th. 2015 (to appear). "Improving Burrows' Delta. An empirical evaluation of text distance measures." In: Digital Humanities Conference 2015.

Hoover, D. 2004a. "Testing Burrows' Delta." *Literary and Linguistic Computing* 19(4), 453–475.

Hoover, D. 2004b. "Delta Prime?" *Literary and Linguistic Computing* 19(4), 477–495.

Koppel, M., Schler, J. and Argamon, S. 2009. "Computational methods in authorship attribution." *Journal of the American Society for Information Science and Technology* 60(1), 9–26.

Oakes, M. P. 2009. "Corpus linguistics and stylometry." In A. Lüdeling & M. Kytö (eds.) *Corpus Linguistics: An International Handbook*, Berlin: Mouton de Gruyter, Berlin, pp. 1070–1090.

Smith, P. and Aldridge W. 2011. "Improving authorship attribution. Optimizing Burrows' Delta method." *Journal of Quantitative Linguistics* 18(1), 63–88.

Stamatatos, E. 2009. "A survey of modern authorship attribution methods." *Journal of the American Society for Information Science and Technology* 60(3), 538–556.

Collocations across languages: evidence from interpreting and translation

Adriano Ferraresi
University of Bologna
adriano.ferraresi@unibo.it

Silvia Bernardini
University of Bologna
silvia.bernardini@unibo.it

Maja Miličević
University of Belgrade
m.milicevic@fil.bg.ac.rs

1 Introduction

Starting with the work of Palmer (1933), interest in the study of collocations has been fuelled by considerations about the difficulties they pose to ESL/EFL learners; a large body of research has since investigated learners' use of collocations (e.g. Nesselhauf 2005; Durrant and Schmitt 2009). More recently, researchers have started to also focus on the use of collocations in what might be seen as the middle ground between native and learner varieties of language, i.e. English as a Lingua Franca and translation. The intriguing suggestion has been put forward that learner, Lingua Franca and translational language should in fact be seen as different forms of constrained communication in language contact situations (Lanstyák and Heltai 2012).

The present contribution aims to compare use of collocations by translators (translating written texts in writing) and interpreters (interpreting spoken discourse orally). We thus hope to provide data of interest not just to corpus-based translation studies (CBTS) and interpreting studies (CBIS) scholars, but in general to all those with an interest in phraseology.

2 Background: collocations in interpreting/translation

Research in CBTS and CBIS has typically adopted the *interlingual parallel* or the *monolingual comparable* approach. Within the former, translated/interpreted target texts (TTs) are compared to their source texts (STs) with the aim to investigate the results of the translator/interpreter's decision-making processes; within the latter, translated/interpreted production is contrasted with comparable original production in the same language, searching for regularities characterizing translated/interpreted language viewed as specific language varieties.

Kenny (2001) and Marco (2009) use interlingual parallel corpora (English/German and English/Catalan respectively) to search for and compare a small number of pre-selected collocations in written STs and the corresponding translated TTs. Jantunen (2004) and Dayrell (2007) adopt a monolingual comparable approach to compare translated and non-translated literary production, focusing on Finnish degree modifiers and high-frequency words in Brazilian Portuguese respectively. Within interpreting studies, collocations have been touched upon in studies of *anticipation* – the strategy whereby interpreters produce a string of words simultaneously with the corresponding string of words uttered by the speaker (Vandepitte 2001).

To the best of our knowledge, neither an in-depth corpus-based study of phraseology in interpreted language, nor a comparison of phraseological patterns in translation vs. interpreting have been carried out so far. For these purposes an *intermodal* corpus is required, featuring parallel or comparable outputs of translation and interpreting tasks – a rather novel corpus set-up (but see Shlesinger and Ordan 2012; Kajzer-Wietrzny 2012).

3 Corpus description: EPTIC

Authentic settings in which translation and interpreting occur are rare, and so are intermodal corpora. This makes the corpus used in this study, i.e. EPTIC (European Parliament Translation and Interpreting Corpus), an especially valuable resource (Bernardini et al. provisionally accepted). The proceedings of the European Parliament are a well-known and widely used source of multilingual texts for NLP applications (see e.g. Koehn 2005). These texts are in fact not the original speeches as delivered at the Parliament, but edited written versions.

In EPTIC we provide, alongside these edited multilingual versions, the transcriptions of the original speeches and of their interpreted versions, as they were initially delivered. Considering all its subcorpora, comprising simultaneous interpretations paired with their STs in Italian and English and corresponding translations and STs, EPTIC is a bilingual (English/Italian), intermodal (oral/written), twice-comparable (original/translated, original/interpreted) and parallel (source/target) corpus.

At the time of writing, the corpus consists of 568 texts, totalling around 250,000 words. It is part-of-speech tagged and lemmatized using the TreeTagger, and indexed with the Corpus WorkBench. Each text is aligned at sentence level with its ST/TT and with the corresponding text in the other modality (oral/written). Metadata encoded with the corpus include speech, speaker and interpreter details (delivery type (read, impromptu, mixed), topic, political party, gender).

4 Aims of the study

The present study aims to assess whether and to what extent English and Italian translators and interpreters in EPTIC differ in their choices to either reproduce a collocation observed in the ST or insert a new one. By applying mixed-effects regression models (Gries to appear), we test the effect on collocation production (reproduction vs. insertion) of mediation mode (translation/interpreting), language direction (English=>Italian and Italian=>English) and collocation association strength, as measured by two lexical association measures that are known to identify frequent vs. strongly-associated word pairs. In so doing, we not only aim to provide quantitative evidence that may confirm or disconfirm the largely qualitative/anecdotal observations made so far on collocations in CBTS and CBIS, but also experiment on the applicability of mixed-effects models to translation and interpreting data.

5 Method

Collocation candidates are extracted from the four target language corpora (interpreted English, translated English, interpreted Italian, translated Italian) based on these patterns:

- adjective + noun (and noun + adjective for Italian): e.g. *political role*, *problema grave* ("serious problem");
- noun + noun: e.g. *road safety*, *autorità (di) bilancio* ("budgetary authority");
- verb + noun: e.g. *exploit (the) dimension*, *ricevono fondi* ("receive funds");
- noun + verb: e.g. *challenges arising*, *passeggeri volano* ("passengers fly").

We discard bigrams including proper nouns, semi-determiners (e.g. *same*, *other*, *former*) and numerals. To evaluate the collocation status of the remaining pairs, frequency data are obtained from ukWaC and itWaC (Baroni et al. 2009) and used to calculate word association strength relying on t-score (*t*) and Mutual Information (*MI*). The cut-off point between collocations and non-collocations is based on the median of the AM scores obtained for the EPTIC-derived bigrams in each language: $MI \geqslant 3$ in English and 5 in Italian, and/or $t \geqslant 6$ in English and 11 in Italian. Bigrams with frequency <3 are excluded (cf. Evert 2008).

Random sets of 150 collocations per corpus are selected for manual analysis. Aligned concordances (ST/TT) are examined to check whether the target text collocation was reproduced or inserted by the translators/interpreter.

The target text collocation status (reproduced/inserted) is used as a binary outcome variable in a mixed-effects logistic regression model with language direction, mediation mode and association measure status as categorical predictors (fixed effects), controlling for possible influences of individual texts, part-of-speech patterns and collocations (random effects). The analysis is conducted using the R package *lme4* (Bates 2005).

6 Results

Texts interpreted into English are found to contain more inserted collocations than the respective translated texts, while the opposite is true for the English=>Italian direction (Figure 1). Overall, the English targets contain more insertions than the Italian targets. In both language directions and both mediation modes, the collocations having both a high *MI* and a high *t* are less likely to be insertions than those scoring high on a single AM.

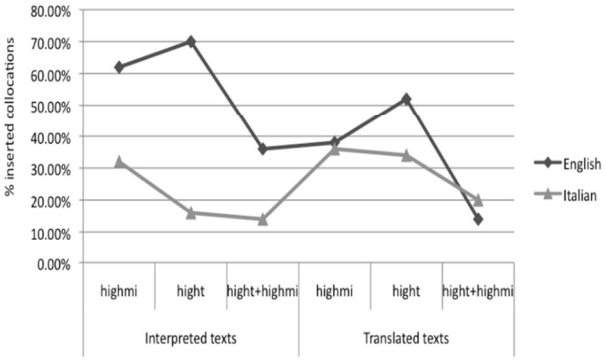


Figure 1. Percentages of inserted collocations by target language, mediation mode and AMs.

All predictors are found to contribute significantly to collocation reproduction/insertion in target texts; significance is also detected for the interaction between language direction and mediation mode. The model coefficients are shown in Table 1.

Fixed effect	Coeff.	SE	*Z*	*p*
(Intercept)	0.612	0.241	2.539	<.05
Language (Italian)	-1.697	0.307	-5.526	<.001
Mediation mode (translation)	-0.995	0.256	-3.885	<.001
Language (Italian)* Mediation mode (translation)	1.505	0.384	3.917	<.001
AM status (high*t*)	0.069	0.221	0.312	ns
AM status (high*t*+high*mi*)	-1.128	0.246	-4.582	<.001

Table1. Summary of the mixed effects logistic regression model.

7 Conclusion and further steps

In this study we have investigated collocations in

translation and interpreting, going beyond anecdotal observations of decisions made by single interpreters and translators, and thus laying the grounds for generalizations about their typical behaviours. The picture that emerges is one in which the production of collocations depends on the mediation mode (oral/written), the language direction (into English/into Italian) and the type of collocation (very frequent and/or strongly associated). Similar factors have also been demonstrated to have an effect on the production of collocations by non-native speakers (Ellis et al. 2008). Investigations of collocations along the lines of the present work may thus be especially rewarding in the search for universals of constrained communication at the phraseological level (Lanstyák and Heltai 2012).

As regards the immediate next steps, our results at the moment only concern cases in which a collocation was observed in the TTs that was either inserted afresh, or transferred from the ST. We plan to replicate the procedure for collocations observed in the STs, to see if translators/interpreters are more likely to reproduce or remove (certain types of) collocations. For instance, interpreters working against time may have automatized routines for rendering certain familiar collocations, but may resort to non-collocational renderings for less common, non-routinized cases.

Secondly, as we study concordances we also register whether a shift in meaning occurred between the ST and TT fragment. Finding out if translators/interpreters are more likely to perform such shifts might tell is if these are more conscious vs. more automatic choices.

Finally, the current approach does not provide any product-oriented quantitative evidence about whether interpreted and translated texts overall contain more/less collocations than each other and than comparable non-mediated texts. A monolingual comparable study along these lines would make an ideal complement for the bidirectional parallel perspective that we have been concerned with in this work.

References

Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. 2009. "The Wacky Wide Web: A collection of very large linguistically processed web-crawled corpora". *Language Resources and Evaluation* 43 (3): 209–226.

Bates, D. 2005. "Fitting linear models in R: Using the lme4 package". *R News* 5: 27-30.

Bernardini, S., Ferraresi, A. and Miličević, M. Provisionally accepted. "From EPIC to EPTIC – Exploring simplification in interpreting and translation from an intermodal perspective". *Target*.

Dayrell, C. 2007. "A quantitative approach to compare collocational patterns in translated and non-translated texts". *International Journal of Corpus Linguistics* 12 (3): 375–414.

Durrant, P. and Schmitt, N. 2009. "To what extent do native and non-native writers make use of collocations?". *International Review of Applied Linguistics in Language Teaching* 47 (2): 157–177.

Ellis, N., Simpson-Vlach, R. and Maynard, C. 2008. "Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL". *TESOL Quarterly* 42 (3): 375–396.

Evert, S. (2008). "Corpora and collocations". In A. Lüdeling and M. Kytö (eds.) *Corpus Linguistics*. Volume 2. Berlin, New York: Mouton de Gruyter.

Gries, S.T. To appear. "The most underused statistical method in corpus linguistics: Multi-level (and mixed-effects) models". *Corpora* 10(1).

Jantunen, J.H. 2004. "Untypical patterns in translations". In A. Mauranen and P. Kujamäki (eds.) *Translation universals: Do they exist?* Amsterdam and Philadelphia: John Benjamins.

Kajzer-Wietrzny, M. 2012. *Interpreting universals and interpreting style*. Unpublished PhD thesis, Adam Mickiewicz University.

Kenny, D. (2001). *Lexis and creativity in translation. A corpus-based approach*. Manchester: St. Jerome.

Koehn, P. 2005. "Europarl: A parallel corpus for statistical machine translation". In *Proceedinggs of MT summit*. Volume 5. Phuket, Thailand.

Lanstyák, I. and Heltai, P. 2012. "Universals in language contact and translation". *Across Languages and Cultures* 13 (1): 99-121.

Marco, J. 2009. "Normalisation and the translation of phraseology in the COVALT corpus". *Meta* 54 (4): 842–856.

Nesselhauf, N. 2005. *Collocations in a learner corpus*. Amsterdam and Philadelphia: John Benjamins.

Palmer, H. E. 1933. *Second interim report on English collocations*. Tokyo: IRET.

Shlesinger, M. and Ordan, N. 2012. "More spoken or more translated? Exploring a known unknown of simultaneous interpreting". *Target* 24 (1): 43-60.

Vandepitte, S. 2001. "Anticipation in conference interpreting: A cognitive process". *Revista Alicantina de Estudios Ingleses* 14: 323–335.

Language learning theories underpinning corpus-based pedagogy

Lynne Flowerdew
Flowerdewlynne@gmail.com

1 Introduction

Corpus-based pedagogy is commonly associated with a 'discovery-based', inductive approach to language learning in which learners extrapolate rules on the basis of their scrutiny of the concordance output. Traditional materials, on the other hand, tend to emphasise rule-based learning. The uptake for corpus-driven learning was rather slow initially, a state-of-affairs encapsulated in Leech's (1997: 2) phrase 'trickle down', but is now widely embraced, albeit remaining at the institutional level. In view of the fact that corpus-based pedagogy embodies a different approach to learning to that of more rule-based traditional materials often using invented examples, it would be useful to take stock of key language learning theories considered to underpin this methodology, namely the 'noticing' hypothesis, constructivist learning and Vygotskyan (1934/1986) sociocultural theories. Of note is that it is only in a few accounts in the literature where these are discussed in depth; in many studies they are left implicit. Thus the key aim of this paper is to make explicit these three language learning theories underpinning much corpus pedagogy with reference to various DDL activities and also to examine to what extent these pedagogic initiatives are supported by findings from relevant empirical studies.

2 'Noticing' hypothesis and DDL

The 'noticing' hypothesis discussed at length in second language acquisition (SLA) studies holds that learners' acquisition of linguistic input is more likely to increase if their attention is *consciously* drawn to linguistic features. Schmidt (1990, 2001), the first to propose this hypothesis, maintains that it precedes understanding and is a condition which is necessary for converting input into intake. Schmidt (2010: 724) has also suggested a related hypothesis, 'noticing the gap', i.e. "that in order to overcome errors, learners must make *conscious* comparisons between their own output and target language input". Frequency issues have also been discussed in relation to noticing, particularly by Ellis (2002) and Swain (1998), who has linked noticing to frequency counts of form. The noticing hypothesis clearly underpins many corpus activities, which, by nature of their methodology, tend to belong to the inductive approach. Although objections have been raised to the noticing hypothesis (Truscott 1998), it could be argued that concordance-based tasks requiring students to attend to *recurrent* phrases would seem to be an ideal means for enhancing learners' input via noticing, leading to uptake.

While the inductive approach, the mainstay of DDL, is entirely dependent on noticing, this can be either student-initiated, involving spontaneous noticing by the learner, or teacher-directed to stimulate noticing of certain features, in line with the more 'guided inductive' approach proposed by Johansson (2009). For example, Kennedy and Miceli's (2010) approach to DDL entails two kinds of noticing activities, 'pattern-hunting' and 'pattern-defining', in their proposed apprenticeship training using a 500,000-word corpus of contemporary Italian to aid intermediate-level Italian students with personal writing on everyday topics. Pattern-hunting techniques included browsing through whole texts on the basis of the title and text-type, and scrutinizing frequency lists for common word combinations. The pattern-defining function was used when students did have a specific target pattern in mind to check. Flowerdew's (2012) DDL tasks using a one-million-word freely available corpus of business letters links required students to notice not only the key-word-in-context but also to scrutinize its co-textual environment to infer contexts in which particular speech acts would be used. For example, for the speech act of 'complaining', data from the business letters corpus revealed that the verb 'complain' was used as a follow-up to a previous complaint, often signaled by a time marker, e.g. *We sent an e-mail complaining of the late shipment last week*. Two small-scale empirical studies reported in the literature provide promising evidence that corpus consultation activities requiring students to apply inductive, noticing strategies are beneficial (see Boulton 2011; Gaskell & Cobb 2004).

3 Constructivist Learning and DDL

In essence, constructivism is an educational philosophy which views acquisition of knowledge as a dynamic process in which the learner is in the driving seat. Collentine (2000: 47) argues that giving learners multiple perspectives (e.g. written sources, network of hyperlinks, video) from which to view a targeted phenomenon "increases the likelihood that the phenomenon will become salient to the learner since features lacking salience in one context might be more salient in another". However, constructivism may not be ideal for all students, on account of their learning style preferences, previous learning background etc., and challenges to this learning theory have been raised (see McGroarty 1998).

A constructivist learner approach has been applied in a few corpus-pedagogic initiatives. The

SACODEYL search tool, for use with spoken language corpora covering seven European languages consisting of interviews with teenagers, offers students four different ways of entry to a corpus to match their needs and learning style preferences, i.e. inductive or deductive (Widmann et al. 2011). As regards the teaching of EAP, Bloch's (2009) program for teaching reporting verbs has a user-friendly interface which allows students to search in two modes, either by a specific word or by concept, which leads the student through five prompt categories. Of note is that Chang's (2012) experiment designed to test whether corpus-based tools afford a constructivist environment is one of the few studies to tackle this issue.

4 Vygotskyan sociocultural theories and DDL

As summarized in Swain (2006: 95), Vygotsky argued that "the development and functioning of all higher mental processes (cognition) are mediated, and that language is one of the most important mediating tools of the mind". Cognition is shaped and reshaped through learners interacting through speech, either dialogic or private, to make sense of meaning. Swain (2006) refers to this dialectical process of making meaning as 'languaging', viewed as an integral part of what constitutes learning. Knowledge is thus co-constructed through collaborative dialogue and negotiation with guidance and support mediated by the teacher or student in the form of 'scaffolding'. However, Weissburg (2008) queries how the premise of inner speech, if accepted, can be developed in instructional activities for L2 writing Notwithstanding Weissburg's reservations, Flowerdew, (2008) reports a corpus-informed report-writing module drawing on the tenets of sociocultural theory. Students were divided into groups with weaker students intentionally grouped with more proficient ones to foster collaborative dialogue through 'assisted performance' for formulating searches and discussion of corpus output. By way of support, Huang's (2011) small-scale experiment provides some evidence that corpus consultation mediated by inter- and intra-group dialogues, conceptualized in terms of Swain's 'languaging', benefits students.

5 Conclusion

It is evident from the above that the noticing hypothesis is referred to more frequently than either constructivist learning or sociocultural theory in DDL, which is probably not so surprising given that the inductive approach usually associated with DDL is underpinned by the 'noticing' of rules and patterns. However, these language learning theories are not uncontroversial and there are only a few related empirical studies that have been carried out. While the results from these studies are promising, additional larger-scale studies of a longitudinal nature are needed to give a more in-depth picture of the beneficial effect of corpus-driven learning, underpinned by the language learning theories elaborated in this paper.

References

Bloch, J. 2009. "The design of an online concordancing program for teaching about reporting verbs". *Language Learning and Technology* 13(1): 59-78.

Boulton, A. 2011. "Language awareness and medium-term benefits of corpus consultation". In A. Gimeno Sanz (ed.) *New Trends in Corpus Assisted Language Learning: Working together*, 39-46. Madrid: Macmillan, ELT.

Chang, P. 2012. "Using a stance corpus to learn about effective authorial stance-taking: a textlinguistic approach". *ReCALL* 24(2): 209-236.

Collentine, J. 2000. "Insights into the construction of grammatical knowledge provided by user-behaviour tracking technologies". *Language Learning and Technology* 36: 45-60.

Ellis, N.C. 2002. "Frequency effects in language processing. A review with implications for theories of implicit and explicit language acquisition". *Studies in Second Language Acquisition* 24: 143-188.

Flowerdew, L. 2008. "Corpus linguistics for academic literacies mediated through discussion activities. In D. Belcher and A. Hirvela (eds.) *The Oral/Literate Connection: Perspectives on L2 Speaking, Writing and Other Media Interactions*, 268-287. Ann Arbor, MI: University of Michigan Press.

Flowerdew, L. 2012. "Exploiting a corpus of business letters from a phraseological, functional perspective". *ReCALL* 24(2): 152-168.

Gaskell, D. & Cobb, T. 2004. "Can learners use concordance feedback for writing errors"? *System* 32(3): 301-319.

Huang, L-S 2011. "Language learners as language researchers: the acquisition of English grammar through a corpus-aided discovery learning approach mediated by intra- and interpersonal dialogues". In J. Newman, H. Baayen & S. Rice (eds.) *Corpus-based Studies in Language Use, Language Learning and Language Documentation*, 91-122. Amsterdam: Rodopi.

Johansson, S. 2009. "Some thoughts on corpora and second language acquisition". In K. Aijmer (ed.) *Corpora and Language Teaching*, 33-44. Amsterdam: John Benjamins.

Kennedy, C. & Miceli, T. 2010. "Corpus-assisted creative writing: introducing intermediate Italian students to a corpus as a reference resource". *Language Learning &*

Technology 14(1): 28-44.

Leech, G. 1997. "Teaching and language corpora: a convergence". In A, Wichmann, S. Fligelstone, T. McEnery & G. Knowles (eds.) T*eaching and Language Corpora*, 1-23. London: Longman.

McGroarty, M. 1998. "Constructive and constructivist challenges for applied linguistics". *Language Learning* 48: 591-622.

Schmidt, R. 1990. "The role of consciousness in second language learning". *Applied Linguistics* 11(2): 129-158.

Schmidt, R. 2001. "Attention". In P. Robinson (ed.) *Cognition and Second Language Instruction*, 3-32. Cambridge: Cambridge University Press.

Schmidt, R. 2010. "Attention, awareness and individual differences in language learning". In W. M. Chan, S. Chi, K. N. Cin, J. Istanto, M. Nagami, J. W. Sew, T. Suthiwan, & I. Walker (eds.) *Proceedings of CLaSIC 2010, Singapore*, 721-737. Singapore: National University of Singapore, Centre for Language Studies.

Swain, M. 1998. "Focus on form through conscious reflection". In C. Doughty and J. Williams (eds.) *Focus on Form in Second Language Acquisition*, 64-81. Cambridge: Cambridge University Press.

Swain, M. 2006. "Languaging, agency and collaboration in advanced second language learning". In H. Byrnes (ed.) *Advanced Language Learning: The contributions of Halliday and Vygotsky*, 95-108. London: Continuum.

Truscott, J. 1998. "Noticing in second language acquisition: a critical review". *Second Language Research* 14: 103-135.

Vygotsky, L. 1934/1986. *Thought and Language*. Cambridge, MA: The MIT Press.

Weissberg, R. 2008. "Critiquing the Vygotskyan approach to L2 literacy". In D. Belcher and A. Hirvela (eds.) *The Oral/Literate Connection: Perspectives on L2 Speaking, Writing and Other Media Interactions*, 26-45. Ann Arbor, MI: University of Michigan Press.

Widmann, J., Kohn, K., & Ziai, R. 2011. The SACODEYL search tool – exploiting corpora for language learning purposes. In A. Frankenberg-Garcia, L. Flowerdew and G. Aston (eds.) *New Trends in Corpora and Language Learning*, 167-178. London: Continuum.

Institutional sexism and sexism in institutions: the case of *Ministra* and *Ministro* in Italy

Federica Formato
Lancaster University
federicaformato.ac@gmail.com

1 Introduction

While there is an extensive literature in sexism in the English language, little has been done in Italian (Fusco, 2012; Robustelli, 2012) particularly in the institutional public space where women have recently had increasing access to. Masculine forms used to address, refer to and talk about female politicians in their role as MPs, ministers and chairs are to be seen within the notion of "overt sexist language"[36] (Litosseliti 2006, Mills 2008), namely when discrimination against women is entrenched in the linguistic form, e.g. *chairman* in English, used to refer to both women and men (Baker, 2008; Mills, 2008). In Italian, forms of overt sexism can be investigated in modification of gendered morphemes. There is room to argue that Italian is a gender-inclusive and fair language with specific gender- specific or gender-free morphemes that, most of the time, indicate whether we are referring to, addressing or talking about individual (or group of) women or men. However, a cultural and social symbolism together with stereotypes have contributed to *change the understanding of grammatical rules* and masculine forms, as I show in this paper, are also used for women, specifically in relation to job-titles.

The investigation presented here starts from a corpus-based quantitative analysis of feminine and masculine forms of *Ministr-* used in 3 widely-read and sold printed Italian newspapers, i.e. *Corriere della Sera*, *Il Resto del Carlino* and *La Stampa*. The newspapers article were collected through the database nexus in the period 2012-2014 to cover the Monti technocrat government (3 female Ministers, end of 2011, beginning of 2013), the Letta (7 female Ministers, April 2013- February 2014) and Renzi (7 female Ministers, February 2014-present) political governments. The paper contributes to the literature on language reform and sexist language in traditionally male-inhabited physical and metaphysical (stereotypes, prototypes) spaces such as the institutional public sphere.

[36]Another form of linguistic sexism is "covert sexism", namely when discrimination is found in the content of what is said or written, e.g. when women are called 'doll'.

2 Data

In order to present what form – the masculine or the feminine – of *Ministr-* is mostly used and what this means in term of gender hierarchy and attitudes, I conduct a corpus-based investigation in the attempt to avoid bias and personal usages. For this study, I built a corpus of almost three years – beginning of 2012 till September 2014 (see explanation below) – of three Italian printed newspapers, i.e. *Corriere della Sera*, *Il Resto del Carlino* and *La Stampa*. This choice is driven by two circumstances, one being the numbers of copies sold on the national soil and the other the availability of the articles in electronic format. On the former – the circulation figures – these newspapers appear in first (*Corriere della Sera*, 464428 copies per year), fifth (*La Stampa*, 229659) and seventh (*Il Resto del Carlino*, 123747) positions according to the *Accertamenti Diffusione Stampa*, i.e. a certifying institutions of circulations of publications. I started with the intention of analysing several newspapers appearing in the list of the most-sold ones. However, I was obliged to compromise over the available data to download for the corpus-based analysis. In order to collect data, I used the Nexis database In Table 1, I present the total number articles analysed:

Corriere della Sera	**24102**
Monti	10147
Letta	8079
Renzi	5879
La Stampa	**20443**
Monti	7396
Letta	7331
Renzi	5716
Il Resto del Carlino	**24508**
Monti	11086
Letta	9985
Renzi	3437
Total	**70715**

Table 1 Total number of newspaper articles which contains *Ministro* and *Ministra* divided into newspapers and governments

3 Methods and analytical framework

In this section, I explain how I conducted the search for the occurrences of *Ministro* and *Ministra* and discuss methodological choices. This investigation aims to collect different forms in order to provide an overall picture of what terms (masculine or feminine) are used.

The following list takes into consideration forms that are 'common' (unmarked, in this case the masculine form) and unusual (marked, in this case the feminine), representing a similar understanding of who is apt and who is still considered as interlopers in Italian politics, that is men and women respectively. To these two linguistic categories, I add a further one defined as "semi-marked" (Formato, 2014), i.e. where only one of the elements in the form used undergoes feminization. The queries inform the study of masculine and feminine forms and their subcategories (+punctuation; + name+surname; + name of the ministry)

- *Ministro*+ name of the ministry, (unmarked form) in its known and used declinations, e.g. *Ministro degli Affari Esteri* (Minister of Foreign Affairs) and *Ministro degli Esteri* (Minister of Foreign (affairs).
- *Ministro* + (name) + surname, (unmarked form) e.g. *Ministro (Elsa) Fornero.*
- ==ministro.,;== (unmarked form), in order to collect instances of *ministr*-followed by punctuation.
- *Ministra* (marked form), e.g. *Ministra Lorenzin*
- *Ministro donna* (lady minister) (marked form)
- *Del/al/ La ministro* (of/at the [feminine] minister [masculine])

These queries aim to provide a thorough analysis of forms of *Ministro* and *Ministra* in the three newspapers and in the three governments in power from 2012 to nowadays. With 17 female ministers in a three-year span, I argue there is room for investigating feminine and masculine forms with relation to this specific office. Having established the language issue – the use of masculine and feminine forms for women– and the political circumstance – the increase in the number of female Ministers– the RQs of this investigation are as follows:

1. What grammatically-gendered form of *Ministr-* was most-used to refer to Italian female ministers in three widely-sold printed Italian newspapers in 2012-14?

2. How are *Ministra* and *Ministro* used when referring to Italian female Ministers in the three governments in the three widely-sold Italian newspapers?

4 Results

In the following tables, I show the results of the study. In table 1, I present the absolute frequencies and percentages of unmarked, marked and semi-marked forms in the three Italian governments.

Table 1 indicates that unmarked forms of reference are more widely used than marked and semi-marked ones. In the three governments, the percentages of unmarked forms are extremely high, ranging from 91.38% in the Monti government,

declining in the Letta government (88.94%) and increasing again in the current cabinet (89.72%). The percentages of marked forms, in relation to unmarked forms, range from 8.14% in the Monti cabinet, to 10.82% and 10.00% in the Letta and Renzi ones, respectively. Semi-marked forms like *La ministro* or *ministro donna* are not extensively used in the data, with percentages showing a decrease from the first government (0.46%) to the last one (0.27%). These results show that these Italian newspapers continue using unmarked forms notwithstanding the increasing in the number of women in the governments.

In terms of sub-categories, Figure 1 shows the trend in their use.

	Monti	Letta	Renzi
Unmarked forms			
AF	**4096**	**4964**	**3963**
%	**91.38**	**88.94**	**89.72**
Marked forms			
AF	**365**	**604**	**442**
%	**8.14**	**10.82**	**10.00**
Semi marked forms			
AF	**21**	**20**	**12**
%	**0.46**	**0.35**	**0.27**

Table 1 Raw numbers (AF) and percentages of unmarked, marked and semi-marked forms of reference in the three governments

The form *Ministro* plus punctuation, signalling anaphoric references to possible previous mentions, is the form that is least used throughout the newspapers and across the governments. Conversely, *Ministra* plus punctuation (or zero) is widely used with an unstable trend as far as the newspapers are concerned in the first government and an increase in the following two with slight differences. In terms of the form *Ministr-* plus name plus surname, the unmarked *ministro* is used more than the marked *ministra*, except in the case of *La Stampa*. This publication uses more than any other form, regardless of un/marked forms, the *Ministr-* plus name plus surname (59.13%) in the Monti government (2011-13). While there is a decrease in the Letta government (2013-14), both unmarked and marked form of *Ministr-* plus name plus surname see an increase in the last government (Renzi, 2014-present) with a difference between the highly-used masculine form (62.47% RC, 64.35% CS and 76.64% LS) and the less-used feminine one (30.83% RC, 37.70% CS, 36.43% LS).

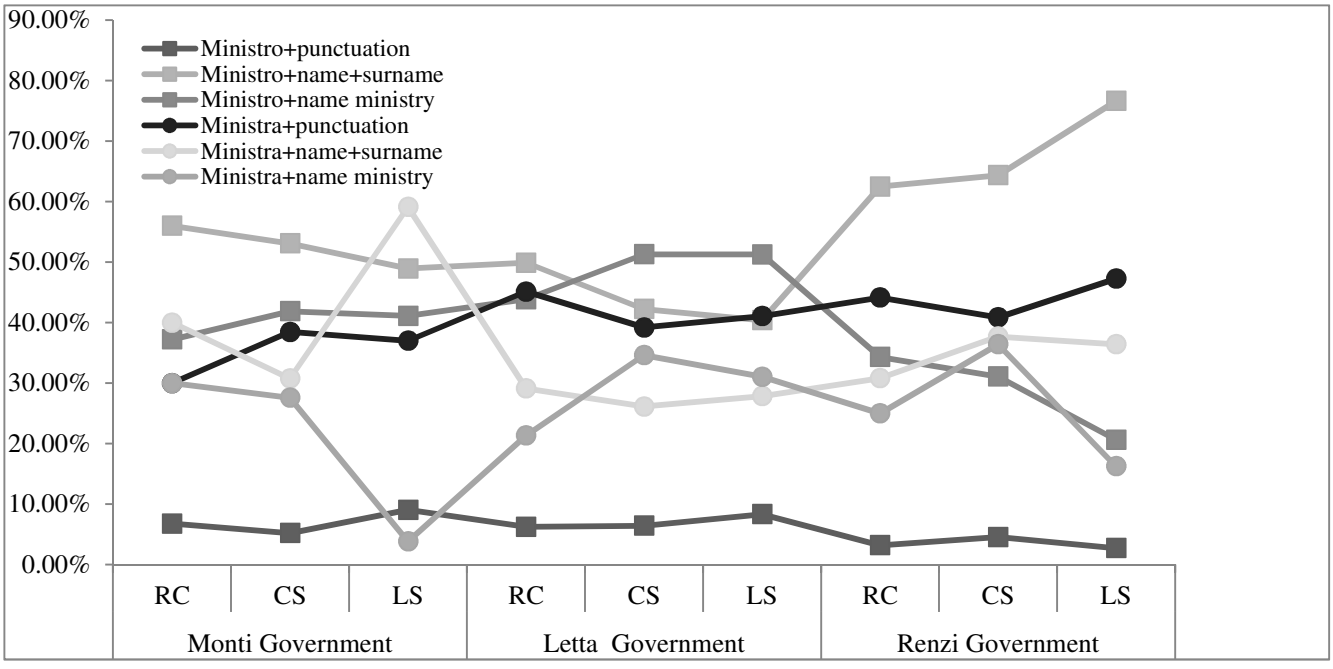


Figure 1 Trends in the use of sub-categories of unmarked and marked forms divided by newspapers and across governments

5 Conclusions

To conclude, while many argue that gendered language determinism might tend to disregard linguistic resistance and habits of speakers, my paper, starting from grammatical accuracy, argues that promotion of a symmetrical linguistic depiction of women and men could be beneficial to gender equality, particularly in male-oriented working spaces.

References

Baker, P. (2008). Sexed Texts: Language, Gender and Sexuality. London: Equinox.

Baker, P. (2014). *Using corpora to analyze gender*. London & New York: Bloomsbury.

Formato, F. (2014) Language use and gender in the Italian Parliament. PhD thesis, Lancaster University. Retrieved from http://www.research.lancs.ac.uk/portal/en/publications/language-use-and-gender-in-the-italian-parliament(12ab6d96-d35e-4062-9628-35036d8fadad).html

Fusco, F. (2012) La lingua e il femminile nella lessicografia italiana. Tra stereotipi e (in)visibilità. Alessandria: Edizioni dell'Orso.

McEnery, T., & Wilson, A. (2001). Corpus Linguistics. Edinburgh: Edinburgh University Press.

Mills, S. (2008). Language and Sexism. Cambridge: Cambridge University Press.

Robustelli, C. (2012). *L'uso del genere femminile nell'italiano contemporaneo: teoria, prassi e proposte*. In Cortellazzo, M. (Ed.), *Politicamente o Linguisticamente Corretto? Maschile e Femminile: Usi Correnti della Denominazione di Cariche e Professioni* (pp. 1-18). Retrieved from http://ec.europa.eu/translation/italian/rei/meetings/documents/decima_giornata_rei_novembre_2010_it.pdf.

Moliere's Raisonneurs: a quantitative study of distinctive linguistic patterns

Francesca Frontini
ILC-CNR
francesca.frontini@ilc.cnr.it

Mohamed Amine Boukhaled
Labex OBVIL
mohamed.boukhaled@lip6.fr

Jean Gabriel Ganascia
LIP6 UPMC
jean-gabriel.ganascia@lip6.fr

1 Introduction and approach

Great authors of plays and novels are often renowned for the ability to create memorable characters that take on a life of their own and become almost as real as living persons to their readers/audience. The study of characterization, that is, of how it is that authors manage to achieve this, has become a well-researched topic in corpus stylistics: for instance (Mahlberg, 2012) attempts to identify typical lexical patterns for memorable characters in the work of Dickens by extracting those lexical bundles that stand out (namely are over-represented) in comparison to a general corpus. In other works, authorship attribution methods are applied to the different characters of a play to identify whether the author has been able to give each of them with a "distinct" voice. For instance (Vogel and Lynch, 2008) compare the dialogues of individual characters in a Shakespeare play against the rest of the play or even against all plays in the Shakespearean corpus.

In Frontini et al (2015), we propose a methodology for the extraction of significant patterns that enables literary critics to verify the degree of characterization of each character with respect to the others and to automatically induce a list of linguistic features that are significant and representative for that character. The proposed methodology relies on sequential data mining for the extraction of linguistic patterns and on correspondence analysis for the comparison of pattern frequencies in each character and the visual representation of such differences.

We chose to apply this analysis to Moliere's plays and the protagonists of those plays. In this work we focus on the figure of the *raisonneurs*, characters who take part in discussions with comical protagonists providing a counterpart to their follies. Such characters were interpreted at times as spokesmen for Moliere himself, and the voice of reason, at other times as comical characters themselves and no less foolish than their opponents. Hawcroft's essay *Reasoning with fools* (2007)

highlights the differences between five of these characters based on their role in the plot. Using this analysis as guidance, we compare significant linguistic patterns in order to see how these differences are marked by the author. We do this by adapting the discourse traits of each of them to the communicative function they needs to fulfill (Biber and Conrad 2009).

2 Syntactic pattern extraction and ranking

In our study, we consider a syntagmatic approach based on a configuration similar to that proposed by (Quiniou et al. 2012). The text is first segmented into a set of sentences, and then each sentence is mapped into a sequence of part of speech (PoS) tags. Tagging is automatically performed using TreeTagger (Schmid, H. 1995)[37]. For example the sentence

J'aime ma maison où j'ai grandi.

is first mapped to a sequence of PoSTagged words;

```
<J' PRO:PER> <aime VER:pres> <ma DET:POS>
<Maison NOM> <où PRO:REL> <j' PRO:PER> <ai
VER:pres> < grandi VER:pper SENT>
```

Then sequential patterns of 3 to 5 elements are extracted. Patterns can be made of PoS Tags only, or of a mix of PoS Tags and recurring lexical elements, with possible gaps (see examples (1), (2), (3)). A minimal filtering is applied, removing patterns with less than 5% of support; nevertheless sequential pattern mining is known to produce (depending on the window and gap size) a large quantity of patterns even on relatively small samples of texts.

In order to identify the most relevant patterns for each of the four characters we used correspondence analysis (CA), which is a multivariate statistical technique developed by (Benzécri, 1977) and used for the analysis of all sorts of data, including textual data (Lebart et al. 1998). CA allows us to represent both Moliere's characters and the (syntactic) patterns on a bi-dimensional space, thus making it visually clear not only which characters are more similar to each other but also which patterns are over/under-represented - that is, more distinctive - for each character or group of characters.

Moreover, patterns can be ranked according to their combined contribution on both axes, and those with the highest contribution can be retained, thus enabling the researcher to filter out less interesting patterns.

3 Analysis and results

CA was performed with the R module *FactoMiner* (Husson et al. 2013) on five characters from five different plays (see Table 1):

Play	**Raisonneur**	**Counterpart**
Ecole des femmes	Chrysalde	Arnolphe
Ecole des maris	Ariste	Sganarelle
Tartuffe	Cléante	Orgon
Mysantrope	Phylinte	Alceste
Malade imaginaire	Béralde	Argan

Table1: Characters and plays.

Figure 1 shows the result of the correspondence analysis, with the five *raisonneurs* printed as blue dots, the patterns printed as red triangles, and the 10 patterns with the highest contribution labeled with their identifiers. Filtering by contribution is crucial in our technique, which extracts over 9500 patterns, most of which are common to all characters (see central cloud in the plot) and thus not so interesting for our study.

The relative distances between the characters seem to match what is already known from literary criticism; first of all Béralde, who is the only character to express himself in prose, is isolated on the right of the x axis. In fact, it is not advisable to compare characters without distinguishing for prose and verse, but we have retained the example of Béralde to show how the proposed technique can easily identify differences in genre.

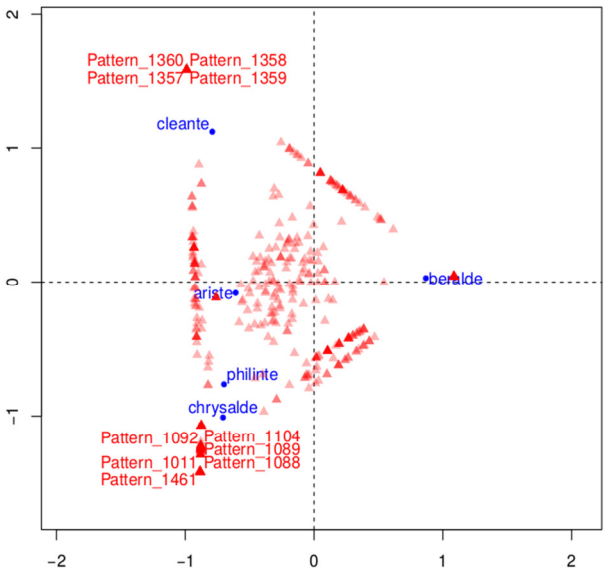


Figure 4 Correspondence analysis plot, with first 10 patterns for contribution

As for the other characters, Hawcroft stresses the difference in the roles of Ariste, Philinte and Chysalde on the one hand and of Cléante on the other. The latter is a more pro-active character, more crucial to the plot; he is also less accommodating than the other three, who are depicted mostly as

[37] For a description of the French tagset see here: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

loyal friends and brothers, trying to help the hero to avoid the consequences of his foolish actions and beliefs.

Instead, Cléante has also to worry about his sister's wellbeing: having to face not only the besotted brother in law, Orgon, but also the man who has duped him, Tartuffe.

In order to confirm this intuition, it is necessary to turn our attention to what it is that exactly causes the spatial distribution, namely the high contribution patterns, we find above. Our technique allows us not only to find the corresponding pattern for each identifier on the plot, but also to extract all underlying instances in the texts. Due to space constraints, only a brief demonstrative analysis will be performed.

Phylinte and Chrysalde are strongly associated with patterns containing prepositional phrases separated by commas. Such patterns are used in contexts where the characters give advice in a very cautious, indirect way. The overuse of punctuation itself, in these two characters, seems to be an indication that the character should be played as a soft spoken person, who is fond of his friend and careful not to offend, e.g.:

(1) Pattern 1011
`[,] [any word] [PRP] [any word] [NOM]`

Instances from **Chrysalde**:

- Entre ces deux partis il en est un honnête **, Où dans l' occasion** l' homme prudent s' arrête [...]
- Il faut jouer d' adresse **, et d' une âme** réduite , Corriger le hasard par la bonne conduite [...]

Instances from **Phylinte**:

- **, Et pour l' amour** de vous , je voudrais , de bon cœur , Avoir trouvé tantôt votre sonnet meilleur .

On the other hand, the patterns most associated with Cléante contain modal constructions, and are indicative of a more direct way of advising, and of stronger arguments, e.g.

(2) Pattern 1360
`[ PRO:PER ] [any word] [VER:infi] [PRP]`

- Les bons et vrais dévots , qu' **on doit suivre à** la trace , Ne sont pas ceux aussi qui font tant de grimace .
- Et s' il **vous faut tomber dans** une extrémité , Péchez plutôt encore de cet autre côté .

Finally, the patterns extracted for Béralde are indicative of the greater simplicity and repetitiveness of his prose, and of the stereotypical role he has in the play, which is that of a man concerned with his brother, as in:

(3) Pattern 865
`[,][DET:POS][any word][PUN]`

- Oui , mon frère , puisqu' il faut parler à cœur ouvert ,

From this experiment it is therefore possible to conclude that the method described above is a promising one, as it not only verifies known facts about the characters in question, but also ground them on corpus based evidence.

4 Preliminary conclusions

Clustering techniques are commonly used in computer aided literary criticism. In order to prove that clusters are significant, statistical analysis can be later applied to verify that resulting clusters are significant. The strength of CA lies in the fact that it allows users to easily identify the reasons for certain texts to group together or to diverge. This helps to overcome the lack of transparency in the presentation of results which often disappoints experts when experimenting with similar techniques, thus making it a useful hermeneutical tool, in the sense of Ramsey (2011)'s algorithmic criticism.

Acknowledgements

This work was supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02, as well as by a scholarship from the *Fondation Maison Sciences de l'Homme*, Paris.

References

Benzécri, J.-P. 1977. Histoire et préhistoire de l'analyse des données. Partie V: l'analyse des correspondances. *Cahiers de L'analyse Des Données*, 2(1), 9–40.

Biber, D., & Conrad, S. 2009. *Register, genre, and style*. Cambridge University Press.

Frontini, F., Boukhaled, M. A., & Ganascia, J. G. 2015 Linguistic Pattern Extraction and Analysis for Classic French Plays. Presentation at the CONSCILA Workshop, Paris.

Hawcroft, M. 2007. *Molière: reasoning with fools*. Oxford University Press.

Husson, F., Josse, J., Le, S., & Mazet, J. 2013. FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R, R package version 1.24.

Lebart, L., Salem, A., & Berry, L. 1998. *Exploring textual data* (Vol. 4). Springer.

Leech, G. N., & Short, M. 2007. *Style in fiction: A linguistic introduction to English fictional*. Pearson Education.

Mahlberg, M. 2012. Corpus stylistics and Dickens's fiction (Vol. 14). Routledge.

Quiniou, S., Cellier, P., Charnois, T., & Legallois, D. 2012. What about sequential data mining techniques to identify linguistic patterns for stylistics? In: Computational Linguistics and Intelligent Text Processing. Springer, (166–177).

Ramsay, S. 2011. Reading machines: Toward an algorithmic criticism. University of Illinois Press.

Schmid, H. 1995. Treetagger| a language independent part-of-speech tagger.Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 43, 28.

Vogel, C., & Lynch, G. 2008. Computational Stylometry: Who's in a Play? In: Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction. Springer, (169–186).

Crawling in the deep: A corpus-based genre analysis of news tickers

Antonio Fruttaldo
University of Naples Federico II
antonio.fruttaldo@unina.it

1 Introduction

Journalistic practices are undergoing, in the last few years, a radical change due to the increasing pressure of new digital media on the professional practice. The ever-growing development of new technologies challenges traditional genres found in this context and, furthermore, analysing genres in a dynamic environment such as that of contemporary journalism calls into question the very nature of genre analysis.

Indeed, genres have been traditionally analysed on the basis of "the use of language in conventionalized communicative settings, which give rise to specific set of communicative goals to specialized disciplinary and social groups, which in turn establish relatively stable structural forms" (Bhatia 1996: 47). On the contrary, in a fluid social context (Deuze 2008), genres are increasingly becoming dynamic rhetorical configurations, whose conventions can be exploited to achieve new goals. In the words of Berkenkotter and Huckin (1995: 6):

> "Genres […] are always sites of contention between stability and change. They are inherently dynamic, constantly (if gradually) changing over time in response to sociocognitive needs of individual users."

Thus, the ultimate aim of genre analysis is becoming that of dynamically explaining the way language users manipulate generic conventions to achieve a variety of complex goals (Bhatia 2004).

Mixed or hybrid forms are most frequently the results of these manipulations, particularly due to the competitive professional environment, where users exploit genre-mixing "to achieve private intentions within the context of socially recognized communicative purposes" (Bhatia 1996: 51). These private intentions, however, are not detectable at first hand, since they are blended in the social context where the hybrid genre is created. Kress (1987) explains this by referring to the so-called "appropriate authority to innovate", which depends on the likelihood of developing new generic forms on the basis of social change. In other words, "unless there is change in the social structures – and in the kinds of social occasions in which texts are produced – the new generic forms are unlikely to succeed" (Kress 1987, 41-42). Thus, if genre-mixing, defined as the "mixture of two or more

communicative purposes through the same generic form (Bhatia 2002: 11), does not meet the appropriate social environment, such forms are less likely to flourish and they will soon perish.

Given the ever-changing social context where journalistic practices operate, they are constantly exploiting new forms of genre-mixing in order to compete with new ways of delivering the news. This intensifying pressure on traditional media has given rise to a variety of mixed-generic forms, among which, in the following paragraphs, we are going to focus on a relatively new genre of TV news broadcast, generally referred to as news tickers (or crawlers).

This genre, which made its first appearance on 9/11 in order to deal with the enormous amount of information coming from the American news agencies, has been adopted by various TV news channels and programmes in order to constantly deliver to viewers a summary of the major news stories of the day or to alert viewers of particular breaking news stories. However, during the years and given the increasing pressure on TV journalism to allure viewers, the genre of news tickers has been slowly appropriating certain generic conventions from other genres to serve this purpose. Indeed, given "[…] the growing ability of viewers to avoid or ignore traditional commercials" (Elliott 2009), TV news networks have found in news tickers a subtle way to market their products, "due to the ticker's location at the bottom of the screen, and its format, which does not interrupt programming" (Coffey and Clearly 2009: 896).

In particular, Coffey and Clearly (2008, 2011) have demonstrated in their work that news tickers can be regarded as overt promotional agents, thanks to the analysis of a corpus of news tickers taken from the American news channels Fox News, CNN and MSNBC. Overt promotional agents are defined by the authors as textual elements that openly advertise the news network itself or its programmes (Coffey and Clearly 2009). Covert promotional agents (or corporate synergy), on the other hand, refers to those textual elements that subtly promote "the parent company's media properties" (Coffey and Clearly 2009: 897). This top-down framework of analysis, however, does not regard other forms of overt promotion that are displayed in news tickers in a subtle way, and that can be highlighted by applying corpus-based methodologies to the analysis of the genre of news tickers.

Thus, in the following paragraphs, thanks to a corpus-based linguistic analysis, we are going to focus on if and how the BBC World News uses its news tickers in order to promote itself and its products. In this, corpus-based methodologies have been of great help, since "The computational analysis of language is often able to reveal patterns of form and use in particular genres […] that are unsuspected by the researcher and difficult to perceive in other ways" (Bhatia 2002: 13). This is the reason why a bottom-up approach to the analysis of these strategies has been adopted, since "one cannot detect these functions without first noticing a pattern of forms" (Berkenkotter and Huckin 1995: 43), which corpus linguistics allows us to do.

2 Collecting and building the NTC Corpus

As previously said, genre analysis is increasingly changing in order to stay up-to-date with the dynamically changing context of contemporary society. This social context has demanded a reshaping of its conventional approach to textual analysis, since genres are progressively becoming fluid entities open to unexpected innovations by borrowing structural conventions and rhetorical configurations from other generic forms. This challenge to genre analysis, however, can be easily overcome by the increasing availability of corpora to researchers. Thus, changes in professional practices can be successfully highlighted by the use of corpus linguistics methodologies.

However, the availability of ready-made corpora may cause some disadvantages on the behalf of the researcher interested in particular areas of human communications, since "a corpus is always designed for a particular purpose" and the majority of them "[…] are created for specific research projects" (Xiao 2008: 383), thus, focusing only on specific genres, while others remain unexplored.

In order to study very specific instances of language in use of a particular discourse community, most of the time, researchers have to create their own specialised corpora, and this is particularly the case of news tickers, given the unavailability of an already-built corpus but, more importantly, no database with instances of this genre.

Thus, the lack of any traces of this genre has forced us to, first and foremost, collect the data by following these steps.

After a week-long preliminary observation and recording of the genre on the BBC World News channel during three parts of the day (i.e., at 8:00 a.m., at 12:00 p.m. and at 8:00 p.m.), in order to decrease the redundancy of news tickers (e.g., news tickers displaying the same information and textual structure) and lower the likelihood of a singular event to dominate the scene, we have decided to focus our attention on the news tickers displayed at 12:00 p.m. during the BBC World News programme *GMT*. We have, then, daily recorded and transcribed in a .txt file thanks to the software Dragon NaturallySpeaking 12.5 (Nuance Communications 2013) the news tickers displayed during this TV

news programme from March 12, 2013 to April 11, 2014 (for a total of 365 days), thus, creating the News Tickers Corpus (NTC), which is comprised of 161,598 tokens (for a total number of 6,937 news tickers). The corpus was, then, annotated through XML encoding, which gives to researchers enough freedom to develop their own coding given the specificities of the genres under investigation (Hardie 2014).

In order to highlight some of the peculiarities found in the NTC corpus, a reference corpus was also collected of all the headlines and lead paragraphs found on the BBC news website thanks to the online database LexisNexis from June 1, 2014 to July 31, 2014. This reference corpus is comprised of 617,311 tokens (for a total number of 20,205 headlines and lead paragraphs) and its selection as a reference corpus was based on the following hypothesis: given the same professional environment, what changes can be highlighted when contents migrate from one textual genre to the other and, more importantly, from one platform to the other. The time discrepancy in collecting the NTC corpus and the reference corpus was also driven by the need to lower the chances that structural similarities were due to identical news contents.

3 Mixing genres and brandcasting the news

The hybrid nature of news tickers is, first and foremost, proved by the merging of two functions traditionally belonging to the journalistic genres of headlines and lead paragraphs. Indeed, while headlines typically "function to frame the event, summarize the story and attract readers", the lead paragraphs work on the information provided in the headline and "describe newsworthy aspects of the event (e.g. the who, the what, the where)" (Bednarek and Caple 2013: 96-97). News tickers, thus, must at the same time catch viewers' attention and give viewers a point of view on the story. However, these two functions coexist with a constellation of other communicative purposes highlighted by structural patterns thanks to the use of the online corpus analysis tool Sketch Engine (Kilgarriff et al. 2004).

One of these communicative purposes can be ascribed to what Meech (1999) defines as brandcasting, which refers to the vast array of corporate branding techniques that broadcasters use in order to project their brand identity. These branding techniques are highly frequent in the NTC corpus and, while some of them may be classified as overt promotional agents (Clearly and Coffey 2008, 2011), others may be seen as subtly achieving the same purpose. In these cases, the authority of the BBC is used in order to legitimise the newsworthiness of the news story found in the news ticker, subtly conveying a subconscious representation in the viewers' mind of the BBC as a source of reliability and trustworthiness. Additionally, these clauses follow a quite strict textual colligation pattern (O'Donnell, Scott, Mahlberg and Hoey 2012) in the textual organization of news tickers, since they are generally placed at the end of the news story reported in the news ticker. In order to see if these brandcasting strategies were not found by chance, we have search for them in the reference corpus and found out that the BBC was rarely used as a source for the news stories, while the name of the reporter was preferred. Thus, these results highlight a difference in the two media and underline how relevant brandcasting is for a TV genre such as that of news tickers, which has found a compromise between its communicative function to inform its viewers/readers and to subtly promote its brand identity.

References

Baker, P. 2014. *Using corpora to analyze gender*. London & New York: Bloomsbury.

Bednarek, M. and Caple, H. 2013. *News discourse*. London & New York: Bloomsbury.

Berkenkotter, C. and Huckin, T.N. 1995. *Genre knowledge in disciplinary communication: Cognition/culture/power*. New Jersey: Lawrence Erlbaum Associates.

Bhatia, V.K. 1996. "Methodological issues in Genre Analysis". *Hermes, Journal of Linguistics* 16: 39-59.

Bhatia, V.K. 2002. "Applied genre analysis: a multi-perspective model". *Ibérica* 4: 3-19.

Bhatia, V.K. 2004. *Worlds of written discourse: A genre-based view*. London: Continuum International.

Bivens, R. 2014. *Digital currents: How technology and the public are shaping TV news*. Toronto: University of Toronto Press.

Bowker, L. and Pearson, J. 2002. *Working with specialized language: A practical guide to using corpora*. London & New York: Routledge.

Coffey, A.J. and Cleary, J. 2008. "Valuing New Media Spaces: Are Cable Network News Crawls Cross-promotional Agents?". *Journalism & Mass Communication Quarterly* 85 (4): 894-912.

Coffey, A.J. and Cleary, J. 2011. "Promotional practices of cable news networks: A comparative analysis of new and traditional spaces". *International Journal on Media Management* 13 (3): 161-176.

Deuze, M. 2008. "The changing context of news work: Liquid journalism and monitorial citizenship". *International Journal of Communication* 2: 848-865.

Fairclough, N. 1992. *Discourse and social change* (16th ed., 2013). Cambridge: Polity Press.

Fairclough, N. 1995. *Media discourse*. London: Hodder Arnold.

Flowerdew, L. 2004. "The argument for using English specialized corpora to understand academic and professional language". In U. Connor and T.A. Upton (eds.) *Discourse in the professions. Perspectives from corpus linguistics*. Amsterdam: John Benjamins, 11-33.

Hardie, A. 2014. "Modest XML for Corpora: Not a standard, but a suggestion". *ICAME Journal* 38 (1): 73-103.

Hoey, M. 2005. *Lexical priming: A new theory of words and language*. London & New York: Routledge.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. 2004. "The Sketch Engine". In G. Williams and S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress: EURALEX 2004*. Lorient: Université de Bretagne-Sud, 105-16.

Kress, G. 1987. "Genre in a social theory of language: A reply to John Dixon". In I. Reid (ed.) *The place of genre in learning: Current debates*. Geelong, Australia: Deakin University Press, 35-45.

Lee, D.Y. 2008. "Corpora and discourse analysis". In V.K. Bhatia, J. Flowerdew and R.H. Jones (eds.) *Advances in discourse studies*. London & New York: Routledge, 86-99.

McEnery, T. and Hardie, A. 2012. *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, T., Xiao, R. and Tono, Y. 2006. *Corpus-based language studies: An advanced resource book*. London: Routledge.

Meech, P. 1999. "Watch this space: the on-air marketing communications of UK television". *International Journal of Advertising* 18: 291-304.

Montgomery, M. 2007. *Discourse of broadcast news: A linguistic approach*. London & New York: Routledge.

Nuance Communications 2013. Dragon NaturallySpeaking (Premium Edition 12.5) [software]. Burlington, MA: Nuance Communications.

O'Donnell, M.B., Scott, M., Mahlberg, M. and Hoey, M. 2012. "Exploring text-initial words, clusters and concgrams in a newspaper corpus". *Corpus Linguistics and Linguistic Theory* 8 (1): 73-101.

Swales, J. 1990. *Genre analysis: English in academic and research settings* (13th ed., 2008). Cambridge: Cambridge University Press.

van Leeuwen, T. 2007. "Legitimation in discourse and communication". *Discourse & Communication* 1 (1): 91-112.

Xiao, Z. 2008. "Well-known and influential corpora". In A. Lüdeling and M. Kyto (eds.) *Corpus Linguistics: An International Handbook*. Berlin: Mouton de Gruyter, 383-457.

Websites

Elliott, S. (2009, January 22). "In 'Trust Me', a Fake Agency Really Promotes. The New York Times. Retrieved September 8, 2014, from http://www.nytimes.com/2009/01/22/business/media/22adco.html

Moore, F. (2001, December 27). "News crawl not just for bulletins anymore". *Pittsburgh Post-Gazette*. Retrieved September 9, 2014, from http://news.google.com/newspapers?id=liQxAAAAIBAJ&sjid=MnADAAAAIBAJ&pg=6570%2C3575355

The truth about news tickers [Web log post] (2011, March 9). Retrieved March 15, 2013, from http://runningheaders.wordpress.com

Poniewozik, J. (2010, November 24). "The tick, tick, tick of the times". *The Time*. Retrieved March 15, 2013, from http://content.time.com/time/specials/packages/article/0,28804,2032304_2032745_2032850,00.html

Wikipedia entry dedicated to News Tickers (2004, September 8). Retrieved March 7, 2013, from http://en.wikipedia.org/wiki/News_ticker

Learners' use of modal verbs with the extrinsic meanings "possibility" and "prediction"

Kazuko Fujimoto
Soka University
kazuko@soka.ac.jp

1 Introduction

Folse (2009: 231) states, "Modals are important to our ELLs [English language learners] because modals help 'soften' a person's speech." He also continues, "Sometimes native speakers perceive the tone of our ELLs as rude or overtly aggressive, and this is often because our ELLs don't use modals." Lea et al. (2014: AWT3) describes, "In academic writing, it is important to use tentative language when making claims", giving modal verbs such as *could*, *may*, *might* for language examples to do so. Modal verbs will help learners to avoid sounding direct in communication, and to express their degree of certainty in academic writing. The aim of this study is to examine Japanese university students' use of modal verbs, comparing with native speaker students' use of them. My corpus-based findings show that the frequency of some modal verbs with extrinsic meanings,[38] which mark tentativeness, is significantly different between the Japanese students and native speaker students. This paper suggests the importance of classroom instruction on how modal verbs are used in the English modality system so that learners can express their attitudes and stance more effectively in English.

2 Methodology

Three corpora were used in this study. The first is the longitudinal one of about 100,000-word written English by 87 second-year Japanese university students who took an academic writing course in the department of the English language in 2009, 2010 and 2012 (Fujimoto Corpus [hereafter FC]). All the students used an academic writing textbook published by Macmillan (Zemach and Islam 2005, 2011[39]), and each student submitted his or her writing assignments every two weeks, eleven times in total per year. The average of the students' TOEIC-IP scores is 458.9 (Range: 210-755; Median: 457.5; SD: 125.4).[40] I also built a rather small textbook corpus of about 6,000 words, which contains written data from sample paragraphs, exercise sections and language notes in Zemach and Islam (2011) (Mac Text Corpus). The third corpus is LOCNESS (Louvain Corpus of Native English Essays), about 300,000-word data from essays by British and American students.[41] Using FC enables the author to identify each student's English proficiency level, and to analyze the students' language use according to the topics for their writing.

First, I compared FC and Mac Text Corpus, and then compared FC and LOCNESS to find the difference or similarity in frequency of modal verbs. All the data were analysed with the computer software AntConc.[42]

Nine central modal verbs *can*, *could*, *may*, *might*, *shall*, *should*, *will*, *would*, *must*[43] and their negative forms including their contracted forms were examined in these corpora. Each modal verb represents all its forms (e.g. *could* represents *could*, *could not* and *couldn't*).

3 Results and Discussion

The frequency of all the modal verbs except *shall* in FC increased steadily as the students submitted their writing assignments about the topics given by the textbook. It could be said that once the students used modal verbs, they started getting used to using them. It is noticed that the frequency of some modal verbs rose up sharply in some of their writing, which seems to be due to the topics about which they wrote. The textbooks do not provide the central modal verbs for useful language expressions for each type of paragraph except *would* in *would like to*.

The difference in frequency of each modal verb between FC and Mac Text Corpus, and between FC and LOCNESS was examined by log-likelihood tests (see Tables 1 and 2). The frequency of *may* is higher in FC than in Mac Text Corpus (the difference is statistically significant at the level of $p<0.01$). The modal verb *would* is more frequently used in Mac Text Corpus than in FC (the frequency difference is statistically significant at the level of

[38] Biber et al. (1999: 485) categorize modal verbs into two types according to their meanings: "intrinsic" and "extrinsic". These two types are also called "deontic" and "epistemic" respectively.
[39] Zemach and Islam (2011) is a new edition of Zemach and Islam (2005), and the content is much the same with some descriptions updated. The former was used for the students in 2009 and 2010, and the latter, for those in 2012.
[40] When the students' TOEIC-IP score range is considered, Zemach and Islam (2005, 2011) were not completely appropriate textbooks to those students. These textbooks were chosen since Academic Writing was a required subject for the second-year students at the author's university, and its purpose was writing paragraphs.
[41] LOCNESS (Louvain Corpus of Native English Essays) was composed by the Centre for English Corpus Linguistics (Université catholique de Louvain, Belgium): http://www.uclouvain.be/en-cecl-locness.html.
[42] Anthony, L. 2014. AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available online at http://www.antlab.sci.waseda.ac.jp/
[43] See Biber et al. (1999: 483).

p<0.01). This is because the frequency of *would* in *would like to* is high in Mac Text Corpus. The frequency of *may* and *would* is significantly much higher in LOCNESS than in FC (at the level of p<0.0001).

	FC	Mac Text Corpus		
	RF	RF	LL	*p*-value
can	464	23	1.62	
could	208	14	-0.03	
may	56	0	6.98	< 0.01
might	31	3	-0.41	
shall	2	0	0.25	
should	187	12	0	
will	252	13	0.72	
would	143	21	-10.17	< 0.01
must	80	0	9.98	< 0.01

Table 1: FC vs. Mac Text Corpus

RF=raw frequency; LL=log likelihood values. Negative values of LL indicate that the modal verb is more frequent in Mac Text Corpus than in FC. Each modal verb represents all its forms (e.g. *could* represents *could*, *could not* and *couldn't*).

	FC	LOCNESS		
	RF	RF	LL	*p*-value
can	464	1321	13.16	< 0.001
could	208	675	0.78	
may	56	475	-50.05	< 0.0001
might	31	83	1.49	
shall	2	11	-0.39	
should	187	796	-6.37	< 0.05
will	252	1102	-11.17	< 0.001
would	143	1510	-218.25	< 0.0001
must	80	316	-1.04	

Table 2: FC vs. LOCNESS

RF=raw frequency; LL=log likelihood values. Negative values of LL indicate that the modal verb is more frequent in LOCNESS than in FC. Each modal verb represents all its forms (e.g. *could* represents *could*, *could not* and *couldn't*).

I focused on the four modal verbs *may*, *might*, *could* and *would* that have the extrinsic meaning "possibility" or "prediction", which is related to tentativeness. The percentage of the students' use of the extrinsic meanings was examined about these four modal verbs in FC. 98.2% of the examples of *may* was used with the meaning of extrinsic-possibility. *Might* was used 100% with the extrinsic-possibility meaning. Only 4.8% of the examples of *could* was used with the extrinsic-possibility meaning. 20.3% of the examples of *would* was used with the meaning of extrinsic-prediction. It should also be added that 61.5% of the examples of *would* was used in the expression *would* [*'d*] *like to* and its negative and questions forms. The high frequency of this fixed phrase in FC may be the influence of the textbook, since it suggests using *would like to* to express wishes, hopes and plans.

The English proficiency levels of the students who used the four extrinsic modal verbs were also examined. The intermediate level students used extrinsic *may* and *might* most, 40.0% and 38.7% respectively. [44] 80.0% of the examples of extrinsic *could* was used by the intermediate and upper intermediate level students, 40.0% for each. Almost half of the examples of extrinsic *would* were used by the upper intermediate level students. Basic level students did not use *might*, *could* or *would* with the extrinsic meanings, and only 1.9% of the examples of extrinsic *may* were from the basic level students.

The corpus data analysis results also showed that the students most frequently used these four extrinsic modal verbs to write a paragraph to express their opinions about a topic assigned by the textbook.

4 Conclusion

This paper focused on the students' use of modal verbs with the meanings extrinsic-possibility and extrinsic-prediction. The corpus analysis results show that the students were likely to use *may* and *might* with the extrinsic meaning, but they used *could* and *would* with the extrinsic meanings much less frequently. Based on this study, I would say that it is necessary to teach the use of tentativeness in academic writing, according to the students' degree of certainty about their claims.

Rundell et al. (2007: IW17) state, "When learners express degrees of possibility or certainty, they often limit themselves to modal auxiliaries, and neglect the many other expressions that can be used for the same purpose." Further research should be conducted to investigate that this can be said about Japanese students as well. It would also be necessary to examine the variation of the students' language use for logical possibility and prediction.

Acknowledgements

I am deeply grateful to Professor Geoffrey Leech for his valuable comments, suggestions and warm encouragement. I pray for my great mentor Professor Geoffrey Leech's eternal happiness and peacefulness. I am also very grateful to Professor Willem Hollmann for his helpful comments and suggestions. All errors and inadequacies are my own.

I would also like to express my deep gratitude to Professor Sylviane Granger for her kind permission to use Louvain Corpus of Native English Essays

[44] The students' TOEIC-IP scores were categorized into the following four levels based on the institutional standard: Basic, Elementary, Intermediate, Upper intermediate and Advanced.

(LOCNESS). This work was supported by JSPS KAKENHI Grant Number 25370654.

References

Anthony, L. 2014. AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available online at http://www.antlab.sci.waseda.ac.jp/

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson Education Limited.

Celce-Murcia, M. and Larsen-Freeman, D. 1999. *The grammar book*. 2nd ed. Boston: Heinle and Heinle.

Coates, J. 1983. *The semantics of the modal auxiliaries*. Oxford: Routledge.

Collins, P. 2009. *Modals and quasi-modals in English*. Amsterdam: Rodopi.

Downing, A and Locke, P. 2002. *A university course in English gramma*r. Oxford: Routledge.

Folse, K. S. 2009. *Keys to teaching grammar to English language learners*. Ann Arbor: University of Michigan Press.

Hyland, K. 2005. *Metadiscourse*. London: Continuum.

Lea, D. (ed.) 2014. *Oxford leaner's dictionary of academic English*. Oxford: Oxford University Press.

Leech, G. 2004. *Meaning and the English verb*. 3rd ed. Harlow: Pearson Education Limited.

Leech, G, Hundt, M., Mair, C. and Smith, N. 2009. *Change in contemporary English: a grammatical study*. Cambridge: Cambridge University Press.

Leech, G. and Svartvik, J. 2002. *A communicative grammar of English*. 3rd ed. Harlow: Pearson Education Limited.

McGrath, I. 2013. *Teaching materials and the roles of EFL/ESL teachers.* London: Bloomsbury.

Mishan, F. and Chambers, A. (eds.) 2010. *Perspectives on language learning materials development*. Oxford: Peter Lang.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. *A comprehensive grammar of the English language*. Harlow: Longman Group Limited.

Rundell, M. (ed.) 2007. *Macmillan English dictionary for advanced learners*. 2nd ed. Oxford: Macmillan Education.

Swan, M. 2005. *Practical English usage*. 3rd ed. Oxford: Oxford University Press.

Sweetser, E. 1991. *From etymology to pragm*atics: metaphorical and cultural aspects of semantic structure. Cambridge: Cambridge University Press.

Tomlinson, B. (ed.) 2003. *Developing materials for language teaching.* London: Bloomsbury.

Tomlinson, B. (ed.) 2011. *Materials development in language teaching*. 2nd ed. Cambridge: Cambridge University Press.

Tomlinson, B. (ed.) 2013. Applied linguistics and materials development. London: Bloomsbury.

Tyler, A. 2012. Cognitive linguistics and second language learning: theoretical basics and experimental Evidence. New York: Routledge.

Zemach, D.E. and Islam, C. 2005. Paragraph writing. Oxford: Macmillan Education.

Zemach, D.E. and Islam, C. 2011. Writing paragraphs. Oxford: Macmillan Education.

Zemach, D.E. and Ghulldu, L.A. 2011. Writing essays. Oxford: Macmillan Education.

Zemach, D.E., Broudy, D. and Valvona, C. 2011. Writing research papers. Oxford: Macmillan Education.

A corpus-based study of English and Thai spatial and temporal prepositions

Kokitboon Fukham
Mahasarakham University
fukham@yahoo.com

Space and time have sparked enormous speculation in science-oriented fields. It is, in fact, that space and time in language are also embodied in language. Space and time in language are reflected through mental representation of such entities as prepositions. Some languages juxtapose concepts of space and time within one preposition and such a preposition entails both temporal and spatial perspectives. In this regard, it can exhibit a concept of time of a particular event and it can also represent a concept of a certain object in relation to a location. Although three English prepositions (*in*, *on* and *at*) and three Thai prepositions (*naj*, *bon*, and *thii*) represent both spatial scenes and temporal frames, they are mentally represented different aspects of space and time concepts. These three prepositions of English and Thai are selected due to their similar surface ideas but internally they are conceptualized differently. The two focal aims of this study are 1) to study mental representation of temporal and spatial prepositions in English and Thai, 2) to examine similarities and differences of space and time through the use of English and Thai prepositions. Trajector-Landmark (TR-LM) framework, used in Cognitive Grammar to demystify mental representation of language structure and use, is adopted as the framework for this study to uncover mental representation as well as differences and similarities between English and Thai prepositions, using a corpus analysis from both languages. The tentative results yield that both English and Thai prepositions pose different dimensions and movements which stem from culture-bound concept of language but they also have some similar features. The findings also imply Thai EFL learners of English should be aware of spatial and temporal differences and if possible, they may have a tendency to master such English prepositions.

Stance-taking in spoken learner English: The effect of speaker role

Dana Gablasova
Lancaster University
d.gablasova@lancaster.ac.uk

Vaclav Brezina
Lancaster University
v.brezina@lancaster.acuk

1 Introduction

Epistemic stance-taking is an important aspect of communicative skills, whether in one's native or non-native language. It plays an essential role in conveying the epistemic perspective of the speaker (i.e. his or her certainty-related evaluation of what is said) as well as in managing and negotiating interpersonal relationships between speakers (Kärkkäinen 2006; Kärkkäinen 2003; Hunston and Thompson 2000). However, despite the significance of stance-taking in everyday discourse (Biber et al. 1999), so far there has been only a limited number of studies that address this issue in second language spoken production (e.g. Aijmer 2004; Fung and Carter 2007; Mortensen 2012). This study therefore aims to contribute to our understanding of this area by exploring how epistemic stance is expressed in the context of a spoken English exam by two groups of speakers – the (exam) candidates (advanced L2 speakers of English) and examiners (L1 speakers of English). In particular, we asked the following two questions:

RQ 1: Is there a difference between the number of certainty and uncertainty epistemic adverbial markers (AEMs) used by the two groups of speakers across different tasks?

RQ 2: Is there a difference between the type of certainty expressed by the two groups of speakers across different speaking tasks?.

2 Method

To answer the research questions, data were taken from a new, growing corpus of L2 spoken production - the Trinity Lancaster Corpus (TLC). The corpus is based on examinations of spoken English conducted by the Trinity College London, a major international examination board, and contains interactions between exam candidates (L2 speakers of English) and examiners (L1 speakers of English). The corpus represents semi-formal institutional speech, and thus complements other corpora of L2 spoken language that may elicit a more informal spoken production (e.g. LINDSEI). In this study, we used the advanced subsection of the TLC which at present contains approximately 0.45 million

words, with almost 300,000 tokens produced by the candidates and about 150,000 tokens produced by the examiners.

The data in this study come from 132 candidates and 66 examiners (some examiners participated in more examinations). The examinations took place in six countries – 31 were conducted in Italy, 31 in Mexico, 30 in Spain, 23 in China, 13 in Sri Lanka and 4 in India. The candidates in the corpus were advanced speakers of English, their proficiency corresponding to C1 and C2 levels of the Common European Framework of Reference for Languages (CEFR). Speech from each candidate was elicited in four speaking tasks – one monologic and three dialogic tasks. All three dialogic tasks are semi-formal in nature and highly interactive. Each sub-component of the exam lasts for about 5 minutes and altogether the corpus contains about 20 minutes of speech from each candidate at the C1/C2 level. Since the exam allows the candidates to bring in their own topics for the *presentation* and some aspects of this topic are also discussed in the *discussion,* the corpus contains spoken L2 production on a great variety of topics. A more detailed description of the exam and each speaking task can be found in the Exam Syllabus by Trinity College London (2010).

3 Procedure

The approach we have chosen was to combine automatic corpus searches with manual analysis to ensure high quality of the results. First, a list of candidate adverbial epistemic markers (AEMs) was compiled based on previous studies that focused on epistemicity, i.e. Holmes (1988), Biber et al. (1999) and Brezina (2012). On this list, there were several forms that are often used also for other than epistemic functions. All of the expressions from this list were searched in the corpus and decisions to exclude some of the words from the list were made on the basis of their primarily non-epistemic functions. In order to answer the second research question, the AEMs signaling certainty were selected and manually coded for different types of certainty identified by grounded analysis.

4 Results and discussion

RQ1: Across all compared tasks the candidates used on average more markers of uncertainty than the examiners with the difference being statistically significant in all cases. No statistically significant differences were found between the two groups of speakers with respect to certainty.

RQ2: Different types of certainty employed by speakers were identified in the data. In particular, three types of contexts in which AEMs of certainty were used: 1. **Subjective use**: In this case, the certainty markers indicate primarily the speaker's positioning towards his or her statement in terms of the degree of certainty. 2. **Intersubjective use**: In this case, while also carrying subjective meaning and expressing a degree of certainty, the epistemic markers are explicitly used to negotiate the speaker's position with respect to the other interlocutor and to react to what he or she has said. 3. Other use: The markers in this category included AEMs whose function could not be clearly categorised as subjective or intersubjective. The results can be seen in Table 1.

As can be seen from Table 1, with respect to the type of certainty expressed, exam candidates performed differently than examiners in the 'interactive task' but in the 'discussion' performed similar to how examiners expressed certainty both in the 'interactive task' and 'discussion'. These findings show that there is no clear-cut difference between how L1 and L2 speakers express certainty; rather L2 speakers modify their epistemic stance-taking according to the interactional setting and their speaker role. The differences between the speaking tasks and the reasons for the candidates' stance-taking choices in each of the tasks will be discussed in the presentation.

Type of certainty	CAND –INT†		CAND –DISC		EX-INT		EX-DISC	
	Freq.	%	Freq.	%	Freq.	%	Freq.	%
Subjective	16	**39.0**	57	**54.8**	35	**68.6**	65	**71.4**
Intersubjective	18	**43.9**	33	**31.7**	10	**19.6**	9	**9.9**
Other	7	**17.1**	14	**13.5**	6	**11.8**	17	**18.7**
Total	41	100	104	100	51	100	91	100

Table 1: Different types of certainty expressed by examiners and candidates
CAND … candidate; EX … examiner; INT … interactive task; DISC … discussion

5 Conclusion

This study sought to demonstrate the effect of different speaker roles and identity on the speakers' linguistic choices when expressing their position (stance) in interaction. We demonstrated that candidates (advanced L2 speakers of English) in an exam differed in their positioning according to the type of speaking task and their role in the interaction which was affected by factors such as familiarity or expertise with the topic discussed and the type of interaction (e.g. discussion of a topic or providing advice to the other speaker). These findings show that when studying L2 spoken production it is important to go beyond characterising the interlocutors as 'native' or 'non-native' speakers of a language. Whereas the fact of being a 'native user' or a 'non-native user' can indeed be part of the speaker role and speaker identity, there are other of equally important factors that arise from the context the exchange. This study thus pointed out the complexity of factors that affect linguistic choices of speakers (whether of L1 or L2), which include the characteristics of the task or interactional setting as well as the role-related expectations and communicative aims.

References

Aijmer, Karin. 2004. Pragmatic markers in spoken interlanguage. *Nordic Journal of English Studies* 3 (1):173-190

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan, and Randolph Quirk. 1999. *Longman grammar of spoken and written English*. London/New York: Longman.

Fung, Loretta, and Ronald Carter. 2007. Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics* 28 (3):410-439.

Hunston, Susan, and Geoffrey Thompson. 2000. Evaluation in Text: Authorial Stance and the Construction of Discourse: Authorial Stance and the Construction of Discourse. Oxford University Press.

Kärkkäinen, Elise. 2003. Epistemic stance in English conversation: A description of its interactional functions, with a focus on I think. Amsterdam: John Benjamins Publishing.

Kärkkäinen, Elise. 2006. Stance taking in conversation: From subjectivity to intersubjectivity. *Text & Talk* 26 (6):699-731.

Mortensen, Janus. 2012. Subjectivity and Intersubjectivity as Aspects of Epistemic Stance Marking. In *Subjectivity in Language and in Discourse*, eds. Nicole Baumgarten, Inke Du Bois, and Juliane House, 229-246. Bingley: Emerald.

Trinity Collecge London. 2010. Graded examinations in spoken English—Syllabus from 1 February 2010.

MDA perspectives on Discipline and Level in the BAWE corpus

Sheena Gardner
Coventry University
sheena.gardner@coventry.ac.uk

Douglas Biber
Northern Arizona University
douglas.biber@nau.edu

Hilary Nesi
Coventry University
h.nesi@coventry.ac.uk

1 Introduction

The design of the BAWE corpus of successful university student writing reflects our assumption that it is worth investigating register variation across levels of study and academic disciplinary groups[45].

	Level 1	Level 2	Level 3	Level 4
Arts & Humanities (AH)	255	229	160	80
Life Sciences (LS)	188	206	120	205
Physical Sciences (PS)	181	154	156	133
Social Sciences (SS)	216	198	170	207

Table 1 BAWE corpus design showing number of assignment texts

The final corpus includes over 6.5 million words from 2,761 successful assignments written by 812 students at four British universities. This paper explores whether evidence of variation across disciplinary groups and years (levels) of study can be found from multidimensional analysis of grammatical features.

2 Multidimensional Analyses of BAWE (dim1988)

Biber has conducted two analyses of the corpus, both of which work on the assumption that an

[45] We also assume variation across genres, and have described the characteristics of thirteen genre families in the corpus (Nesi and Gardner 2013; Gardner and Nesi 2013). The British Academic Written English (BAWE) corpus was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (formerly of the Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800).

academic register can be described in terms of multiple dimensions of variation, each of which emerges from statistical analysis of features that cluster in texts, or are noted by their absence, and that these dimensions can be interpreted as meaningful choices made in context. The first analysis describes the BAWE corpus texts in terms of the 1988 dimensions and enables us to locate academic writing in relation to non-academic genres such as romance fiction and conversation. This indicates that student writing becomes increasingly informational, less narrative, more elaborated, less overtly persuasive and more impersonal from first year undergraduate (level 1) to taught post graduate (level 4).

LEVEL (Year)	More Informational	Less Narrative	More Elaborated	Less overtly Persuasive	More Impersonal
1	−12.7	−2.7	5.1	−1.4	5.9
2	−13.9	−2.8	5.6	−1.4	6.2
3	−14.7	−3.0	5.7	−1.5	6.4
4	−17.2	−3.2	6.3	−2.0	5.5

Table 2. Levels in the BAWE corpus (1988dims)

Disciplinary Group	Informational	Non Narrative	Elaborated	Not overtly Persuasive	Impersonal
AH	−13.4	−2.1	5.7	−2.3	5.5
SS	−15.3	−3.0	6.5	−1.3	6.2
LS	−15.6	−3.0	5.7	−1.5	5.7
PS	−13.4	−3.7	4.4	−1.2	6.5

Table 3. Disciplines in the BAWE corpus (1988dims)

In contrast to the levels which are clearly sequenced across the first four dimensions, the disciplinary groups are characterised by their relative positions: Arts and Humanities are most narrative, Social Sciences most elaborated, and Physical Sciences most impersonal. Such findings are developed in Nesi and Gardner (2012) *inter alia* in the descriptions of genres of assessed student writing.

3 Multidimensional Analyses of BAWE (dim2010)

In 2010 Biber conducted a new factor analysis of the corpus that identified four dimensions specific to student academic writing. As Table 4 shows, student writing in first year undergraduate (Level 1) is at the opposite end of each dimension from writing in fourth year undergraduate or first year postgraduate (Level 4) in all dimensions. These differences are significant for all factors, with only factor 2 showing significant differences between all four levels, as indicated by the superscripts A-D, following DuncanGroup Means tests.

	N texts	Factor 1	Factor 2	Factor 3	Factor 4
Level 1	795	-2.0349 A	1.5277 A	0.9677 A	0.1515 A
Level 2	754	-0.6158 B	0.4411 B	0.3448 BA	0.3219 A
Level 3	589	0.1279 B	-0.484 C	-0.0385 B	0.1356 A
Level 4	598	3.3557 C	-2.1104 D	-1.6833 C	-0.7409 B

Table 4 BAWE 2010 factors by level of study

With reference to the loaded features in each dimension, from D1, we might infer that student writing becomes more dense (more nominalisations, higher TTR, longer words); from D2 we might infer that student writing becomes more technical and concrete (more technical/concrete nouns, fewer human and pronouns); from D3 we might infer that student writing becomes less cognitive and more causative; and from D4 we might infer that student writing increasingly involves fewer timeless truths and more past events.

The standard deviations for these factors range from 3.5 to 9.1, so the range of values is large, nevertheless, the table also suggests that writing at Levels 2 and 3 is quite similar, and positioned between Levels 1 and 4, and closer to Level 1 in factors 3 and 4. This makes sense if we consider that first year students are still finding their academic writing feet, there is gradual progression through years one to three, then a step change to Level 4. This step change is also seen in Durrant's (2013) work on vocabulary overlaps in BAWE, which shows that the vocabulary used in specific disciplines can be similar at levels 1, 2 and 3, but rather different in Level 4. For example, in Engineering the vocabulary at level 4 is more similar to that of management studies which reflects a shift in the focus of study. A further contextual factor that may explain this shift in some areas is that while most undergraduate students represented in the corpus are UK students, the number of international students is greater at level 4. Finally Level 4 can involve a change of discipline for students (e.g. from English BA to MA in Applied Linguistics; from Economics BSc to MBA; from BA in Media and Communication to MA in Publishing).

In Table 5 we see the spread of results for disciplinary group. Here we see writing in the Physical Sciences is at one extreme on all dimensions and can be characterised as activity focused (time adverbs, concrete nouns); Writing in Arts and Humanities is differentiated clearly in

dimensions 2 and 3 which would characterise it as using human participants engaged in sensing and cognition, while dimension 1 differentiates writing in the Social Sciences as specifically dense and theoretical. These characterisations resonate with our knowledge of writing across the disciplinary groups.

	N texts	Factor 1	Factor 2	Factor 3	Factor 4
AH	654	0.8968525 B	2.8767694 A	4.6292360 A	-1.3375604 C
SS	698	4.7346048 A	0.9786615 B	1.6961467 B	0.2192967 B
LS	611	-0.9857905 C	-1.0196053 C	-2.6694000 C	0.3171914 BA
PS	773	-5.8237468 D	-3.1525827 D	-4.0513920 D	0.7918961 A

Table 5 BAWE 2010 factors by Disciplinary Group

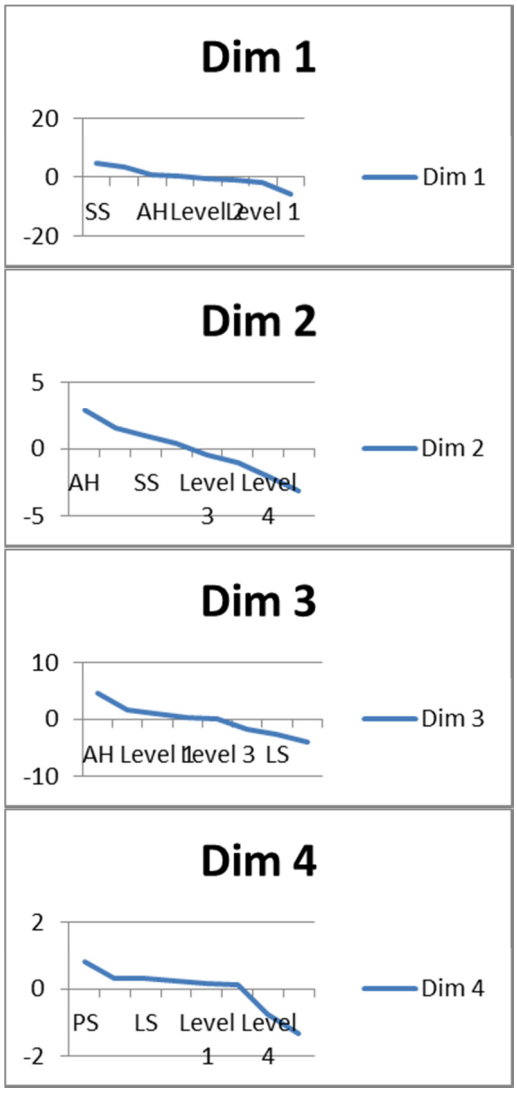


Figure 1 BAWE 2010 Dimensions

In Figure 1 we present the two sets of data together. In Dim 1, SS and L4 contain language that is densely packed with nominalisations and abstract processes, in contrast to PS and L1 which involves activity and concrete nouns. In Dim 2, AH and L1 involve more first person volition and attitude in contrast to the more technical L4 and PS. In Dim 3, AH and SS are characterised by ideas and theories in contrast to LS and PS which involve agents and causes. In Dim 4, PS and L2 are characterised by modality and non-past tense in contrast to the past tenses of history found in AH, and of research reported in L4.

With significant differences between all disciplinary groups on three of the four factors, the differences between disciplinary groups are greater than those between levels in all dimensions. It is also clear that PS is always at one extreme, with LS the next discipline, so the language of writing in the Life and Physical Sciences is grammatically consistently differentiated from writing in Social Sciences and from writing in Arts and Humanities.

In the presentation, the specific features in each dimension and examples of text extracts will illustrate the dimensions, and contrasts with Hardy and Romer (2013)will be discussed.

On one level these may sound like trivial findings, but on others it is a significant contribution to the arguments in favour of teaching EAP students from similar disciplinary backgrounds together. The arguments for a 'common core' of EAP inherent in the pursuit of general academic word lists, and the arguments from EAP and subject tutors that there should be more focus on 'basic' English are difficult to maintain in the face of evidence that the writing demands of the disciplines are so different lexically (Durrant 2013) and as we argue here, grammatically.

References

Biber, D. 2012 Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory* 8(1), 9-37.

Durrant, P. 2013 Discipline and level specificity in university students written vocabulary *Applied Linguistics*

Gardner, S. and H. Nesi (2013) A classification of genre families in university student writing. *Applied Linguistics* 34 (1) 1-29

Hardy, J. and U. Romer 2013 Revealing disciplinary variation in student writing: a multi-dimensional analysis of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora* 8 (2) 183-207.

Nesi, H. and S. Gardner 2012 *Genres across the Disciplines: Student Writing in Higher Education* Cambridge Applied Linguistics Series, CUP

Analysing the RIP corpus: the surprising phraseology of Irish online death notices

Federico Gaspari
University for Foreigners of Reggio Calabria "Dante Alighieri"
gaspari@unistrada.it

1 Motivations and objectives of the study

Although one cannot but agree with Queen Gertrude's statement in *Hamlet* (W. Shakespeare, Act I, scene 2, line 72) that "Thou know'st 'tis common; all that lives must die", it is fair to say that most people are somewhat uncomfortable with the prospect of dying, whether this concerns themselves or their loved ones. Judging from the paucity of research on death-related discourse – in spite of its obvious enduring significance to humankind – linguists, including those of the corpus persuasion, seem to share the general aversion to addressing this universal phenomenon.

Obituaries have arguably received more attention within the funereal genre, not only from scholars of cultural studies, ethnography and sociology (e.g. Fowler 2007), but also, crucially, from linguists – examples include Moore (2002), Al-Khatib and Salem (2011) and Loock and Lefebvre-Scodeller (2014). The fascination with obituaries as the most appealing funereal text type for linguistic inquiry may be due partly to their being about high-profile and often popular (but distant) dead people, and partly to the aesthetic value ascribed to tributes celebrating the remarkable lives of these public figures (Starck 2009) – Moses and Marelli (2003:123) go so far as to claim, possibly overstating their case a little, that "obituaries are perhaps the most frequently read section of the daily newspaper" (see also Starck 2008).

In contrast, death notices (DNs) represent a much more common funereal text type, especially in English-speaking countries, which however has been largely neglected by researchers, with few notable exceptions such as the diachronic accounts by Lipka (2002:62ff) and Fries (2006). DNs typically concern deceased ordinary folk, and serve important social functions like informing those who knew the departed about mourning and funeral plans that they might want to take part in. Once a poor relation of obituaries as much shorter, and paid, newspaper announcements about the recent demise of mostly obscure individuals, DNs have received a new lease of life, so to speak, on the Internet: several specialised websites publish them daily and store old notices in online searchable archives.

The lack of large-scale analyses of web-based DNs is therefore quite surprising for two main reasons: firstly, because the long-term relevance of this sub-genre to humanity seems indisputable; and, secondly, because the discourse of online DNs can offer valuable socio-cultural insights, as has been the case with death-related customs and funereal practices since time immemorial (Petrucci 1995). In an attempt to reverse the long-standing unwillingness of corpus linguists to investigate funereal text types, this study analyses the most frequent lexico-phraseological patterns in a single-source corpus of nearly 240,000 Internet-derived DNs.

2 Construction, composition and size of the RIP corpus

Since DNs typically are rather short texts, many of them must be collected for a corpus-based study to allow for meaningful phraseological and possibly also sociologically-oriented analyses. The RIP corpus was built semi-automatically, crawling the "RIP.ie – End of Life Matters" website[46] to collect all the available DNs, published daily from July 2006 until the end of 2014.

A series of automatic filtering and cleaning routines accompanied by systematic manual checks were applied iteratively to remove all ads, boilerplate, HTML tags and similar interfering material, to obtain a clean, high-quality corpus: Table 1 shows the amount of online DNs with the number of tokens (broken down by year) and the total size of the RIP corpus that resulted from this process, which corresponds to the data analysed for this study.

Year	No. of DNs	No. of tokens
2006 (from July)	10,291	616,692
2007	22,127	837,540
2008	28,620	1,152,719
2009	29,098	1,346,398
2010	28,770	1,158,862
2011	26,918	1,638,904
2012	27,995	1,965,324
2013	31,376	2,575,549
2014	34,566	3,112,375
Total size	*239,761*	*14,404,363 (types: 29,239)*

Table 1: Components and size of the RIP corpus

3 Key features of online death notices

Funeral oratory is a time-honoured tradition in countless language and cultural communities around the world, and most people encounter a wide range

[46] The website, which has fully searchable archives, is available at www.rip.ie (last accessed on June 1st, 2015).

of spoken and written death-related texts on a regular basis, including obituaries, epitaphs, orations, eulogies, elegies, etc. Within this rather heterogeneous group, and despite a certain degree of inevitable variability both in terms of the quantity and of the type of information that they might contain, online DNs are distinctive in that they normally present a well-codified structure, with a fairly predictable combination of these conventional elements: a brief (usually praising) description of the deceased as a tribute to their memory, some details about the circumstances of their passing, the expression of sadness and loss felt by family and friends, ending with the rather mundane practicalities of funeral and burial arrangements, religious and memorial services, etc.

Among the remarkable features of the online DNs under investigation are the frequent contamination with other genres and the mingling of registers: alongside the relatively formal affectionate portrayal of the deceased and factual statements concerning the circumstances of their death, funeral details, etc., one often finds sombre quotes from the Scriptures, highly emotional passages of religious hymns and prayers, liturgical terms, passionate poems or intimate dedications from family and friends, and occasionally farewell formulae and blessings.

4 Phraseological analysis of the Irish online death notices in the RIP corpus

The average length of individual online DNs in the RIP corpus is approximately 60 tokens. While a few of them are very short, vague and uninformative (an extreme example being *Died in London.*), there are others that, in addition to giving detailed information on the funeral arrangements, also dwell at some length on the deceased, describing their family, professional and leisure activities, etc.

A lexical analysis of the RIP corpus reveals that straightforward and common vocabulary that might in principle be used in connection with somebody who has passed away is systematically avoided in online DNs. This is the case, in particular, for the word forms *death* (ranked only 129th in the wordlist, with just 957 occurrences per million words or 0.057 on average per DN, almost always in the sentence-initial standard phrases *The death has occurred of*, *The death has taken place* or *The death took place*), *died* (ranked 339th, 227 pmw, 0.014 pDN), *dead* (4,775th, 4 pmw, 0.0002 pDN) and *body* (1,495th, 30 pmw, 0.002 pDN): these ordinary lexical items turn out to have surprisingly low frequencies in the RIP corpus, due to the intense substitutional competition of euphemistic variants (cf. Fries 1990:60) including *reposing* (21st, 10,104 pmw, 0.6 pDN) and *remains* (175th, 558 pmw, 0.03 pDN) – it is noteworthy that due to register constraints there are no occurrences of *corpse* or *cadaver* in the 14.4-million-word RIP corpus.

These results can be put in perspective by looking at the rankings and occurrence rates of some high-frequency words in the RIP corpus, in particular: *funeral* (9th in the wordlist, 20,046 pmw, 1.2 pDN), *church* (11th, 15,802 pmw, 0,95 pDN), *mass* (12th, 15,605 pmw, 0.94 pDN) and *cemetery* (14th, 14,015 pmw, 0.84 pDN).

The analysis also focuses on adverbs used to characterise the circumstances of dying, e.g. *peacefully* (4,262 pmw, 0.26 pDN), *suddenly* (1,176 pmw, 0.07 pDN), *unexpectedly* (427 pmw, 0.02 pDN) and *tragically* (89 pmw, 0.005 pDN). Other phraseological patterns found in online DNs of interest for a sociologically-oriented analysis include the immediate left-hand adjectival collocates and the wider co-text of kinship terms, particularly typical adjectives qualifying bereaved or deceased family members: *beloved husband* (1,239 pmw, 0.07 pDN), *loving husband* (476 pmw, 0.03 pDN), *beloved wife* (1,496 pmw, 0.09 pDN) and *loving wife* (869 pmw, 0.05 pDN).

Finally, the discussion concentrates on 2/3/4-word lexical bundles, revealing the heavily formulaic nature of online DNs in the RIP corpus: the stock of frequently recurring phrases is drawn especially from the domains of culture-specific mourning traditions and religious rituals, e.g.

- *reposing at* (9,006 pmw, 0.54 pDN);
- *funeral home* (7,656 pmw, 0.46 pDN);
- *requiem mass* (5,942 pmw, 0.36 pDN);
- *rest in peace* (5,793 pmw, 0.35 pDN);
- *burial afterwards in* (4,842 pmw, 0.29 pDN);
- *sadly missed by* (3,608 pmw, 0.22 pDN);
- *donations if desired to* (3,016 pmw, 0.18 pDN);
- *family flowers only please* (2,516 pmw, 0.15 pDN);
- *followed by burial in* (2,493 pmw, 0.15 pDN).

5 Conclusions and future work

This study based on a 14.4-million-word corpus of almost 240,000 Irish online DNs has found that they possess a number of interesting lexico-phraseological features, some of which were quite unexpected. The overall findings reveal that this so far neglected web-based funereal text type shows distinctive phraseological, rhetorical and socio-cultural characteristics that deserve further scrutiny, ideally with larger data sets coming from multiple sources (including printed ones), also covering different geographical areas and longer time spans.

This exploration of online DNs in the RIP corpus is part of a larger ongoing research project, which

investigates the specificities of funereal discourse looking at a range of death-related text types in English. This work in progress also involves the construction and analysis of diachronic multi-source corpora of obituaries, funeral orations and eulogies, collecting data from different English-speaking countries.

References

Al-Khatib, M. and Salem, Z. 2011 "Obituary announcements in Jordanian and British newspapers: A cross-cultural overview". *Acta Linguistica* 5 (2): 80-96.

Fowler, B. 2007 *The Obituary as Collective Memory*. Abingdon: Routledge.

Fries, U. 1990. "Two Hundred Years of English Death Notices". In M. Bridges (ed.) *On Strangeness*. Tübingen: Gunter Narr. 57-71.

Fries, U. 2006. "Death Notices: The Birth of a Genre". In R. Facchinetti and M. Rissanen (eds.) *Corpus-based Studies of Diachronic English*. Bern: Peter Lang. 157-170.

Lipka, L. 2002. "Non-serious text types and German death notices – an unlikely pair". In A. Fischer, G. Tottie and P. Schneider (eds.) *Text Types and Corpora: Studies in Honour of Udo Fries*. Tübingen: Gunter Narr. 59-66.

Loock, R. and Lefebvre-Scodeller, C. 2014. "Writing about the Dead: A Corpus-based Study on How to Refer to the Deceased in English vs French Obituaries and Its Consequences for Translation". *Current Trends in Translation Teaching and Learning E* 1 (2014): 115-150.

Moore, S.H. 2002. "Disinterring ideology from a corpus of obituaries: A critical post mortem". *Discourse & society* 13: 495-536.

Moses, R.A. and Marelli, G.D. 2003. "Obituaries and the discursive construction of dying and living". *Texas Linguistic Forum* 47: 123-130.

Petrucci, A. 1995 *Le Scritture Ultime: Ideologia della Morte e Strategie dello Scrivere nella Tradizione Occidentale*. Torino: Giulio Einaudi.

Starck, N. 2008 "Death can make a difference: A comparative study of 'quality quartet' obituary practice". *Journalism Studies* 9 (6): 911-924.

Starck, N. 2009 "Obituaries for sale: Wellspring of cash and unreliable testimony". In B. Franklin (ed.) *The Future of Newspapers*. Abingdon: Routledge. 320-328.

A golden keyword can open any corpus: theoretical and methodological issues in keyword extraction

Federico Gaspari
UniStraDA
gaspari@unistrada.it

Marco Venuti
Univerisity of Catania
mvenuti@unict.it

1 Keyword-related issues in corpus linguistics

A standard technique in corpus linguistics and corpus-assisted discourse studies consists in automatically extracting keywords from corpora, to identify the content words that stand out in terms of frequency and keyness. This is usually a crucial stepping stone for further lexico-phraseological investigations, and the most widely used corpus analysis tools support this function, providing a choice of keyness measures that can be used for keyword extraction and ranking.

One major factor involved in generating a reliable keyword list concerns the reference corpus that is used, primarily with regard to its size and to the number of texts that it contains – not (only) in absolute terms, but (also) relative to the focus corpus (Kilgarriff 2009; Scott 2009). Other criteria also apply, such as the time span of the texts, that should be roughly similar, the number of different authors represented to prevent individual bias, etc. Further serious complications arise for variationist studies, which are the focus of this paper and investigate the differences and similarities between two or more corpora that are somehow related to one another, but typically differ by one major variable. Such studies may compare, for example, the phraseology of news reports on a certain topic published by two different newspapers over the same period of time (with the source being the independent variable, say broadsheet vs. tabloid). To generate the keywords reflecting the relevant independent variable for each focus sub-corpus (for instance: broadsheet newspaper), the researcher might select the "obvious" counterpart of the sub-corpus in question (tabloid, in this case) to act as an ad hoc reference sample; otherwise, one might choose an independent external reference corpus for all cases. Even if the researcher makes such methodological choices carefully, the question remains of how they affect the resulting keywords – and as a consequence the entire study based on them.

An additional problem concerns the distribution of keywords across the texts included in the focus corpus, which is a particularly thorny issue also because of the growing tendency to use large

corpora containing several texts: what dispersion pattern is required for lexical items to qualify as keywords? This is another subtle and complex issue that may escape the direct control and conscious decisions of the researcher, but which is also bound to affect the selection and reliability of the keywords chosen as starting points for any corpus-based research (Baker 2004).

Finally, another important issue is that keywords, by their very nature, tend to overemphasise differences when they are extracted from corpora that are subject to comparison. This is one of the reasons why it has been suggested that it is appropriate to also consider what is relatively similar and stable, in addition to what varies, across corpora. To this end, the identification and analysis of 'lockwords' (Baker 2011; Taylor 2013) are useful steps for more accurate corpus-based lexico-phraseological comparisons.

2 Motivations and objectives of the study

The authors of this paper have themselves confronted these and other similar issues on several occasions, and appreciate the implications of careful keyword selection – we believe that many other members of our community are in a similar situation and share our concerns. While the methodological sections of corpus linguistics papers and major reference works (e.g. Bowker and Pearson 2002; Scott and Tribble 2006) provide helpful starting points that the authors have used for guidance, a systematic and unified treatment of these theoretical and methodological foundations seems overdue. We contend that an inclusive and thorough discussion of these foundational issues is bound to benefit the whole community, also with a view to strengthening the validity of the findings of corpus studies.

Building on the recent work by Gabrielatos and Marchi (2012) and Cvrček and Fidler (2013), this paper wishes to contribute to a timely debate that can at least point to shared best practice in our community. While it would be presumptuous and foolish to give detailed guidelines on keyword extraction that can apply once and for all, there is a need to tackle the relevant methodological assumptions head-on; in this spirit, the paper intends to explore the underlying issues with two case studies, in which the main factors at play are manipulated in different permutations, to draw methodological lessons of wider applicability. We are well aware of the enormous variability of corpus-based studies, depending on the objectives of each project and on the specificities of the data sets involved: by focusing on typical major scenarios, we hope to discuss clearly the most frequent central issues involved in keyword and lockword selection, offering our take on methodological and operational questions of general interest.

3 Related work

Keyword extraction is of utmost importance in corpus studies of different types, regardless of the approach that is adopted. Dedicated collections like Scott and Tribble (2006), Archer (2009) and Bondi and Scott (2010) are characterised by a methodological focus on keyword selection, and the breadth of studies based on preliminary keyword identification is impressive. Among them we can mention by way of example Baron et al. (2009), who explore the techniques of keyword analysis applied to historical corpus linguistics, and Kemppanen (2004), who shows how keywords convey ideology in translated and original language, using translations from Russian into Finnish and non-translated Finnish texts of political history. Even such a cursory overview demonstrates the centrality of keyword extraction for the full range of corpus linguistics research. Hence the need to seriously discuss the theoretical and methodological underpinnings in a critical but unified fashion, without unquestioningly replica ting well-established practice.

4 Between theory and methodology: a critique of standard keyness measures

Starting from the data used in two earlier corpus-based variationist phraseological studies (Gaspari 2013, 2014), we aim to provide an evaluation of previous results by comparing them with those obtained with the approach suggested by Gabrielatos and Marchi (2012).Our main aim is that of testing their approach to keyword identification staring from the results of the previous analysis. In other words we want to show to what extent the new approach contributes to a different interpretation and to a finer-grained analysis of differences and similarities across our corpora.

Our analyses are based on two corpora. The first includes official biographical profiles and award motivations of Nobel Prize winners between 1901 and 2013, while the second consists of maiden speeches (i.e. speeches delivered by new members of the British Parliament when they address the House for the first time) between 1983 and 2011. Each corpus contains around 1.3 million words, with male and female components, representing forms of Institutional Discourse (Drew and Heritage 1992), and displaying features of established genres together with the evaluation of personal achievements (official biographical profiles of Nobel Prize winners and motivations of Nobel Prize awards) and the expression of personal style (maiden speeches).

Our working hypothesis is that through the comparison of male/female corpus components it is possible, and indeed beneficial for accurate investigations, to separate lexico-phraseological features that pertain to the genre (lockwords) from those characterizing gender-related differences in projecting male and female institutional identities that are manifested by corpus-specific keywords. In particular, lockwords in the Nobel Prize Corpus include lexical items used to celebrate the lives and achievements of Nobel Laureates regardless of their gender, such as *successful*, *leading*, *distinguished*, *outstanding*, etc. In contrast, keywords in the female sub-corpus show that women winners are more likely to be remembered for more intimate relations (*husband*, *maternal*, *girl*, *sister*, *children*, *kids*, *families*) and for a less extrovert personality (*contemplative*, *dearly*, *unassuming*). Male winners' personalities and achievements, on the other hand, are described as *competitive*, *rational*, *fruitful*, and *fundamental*.

The analysis of keywords and lockwords in the UK Maiden Speeches Corpus shows similar trends in the way male and female newly elected MPs address their House for the first time. In order to validate what we regard as regularities along the male/female divide, we also compare the gender components of the corpora in order to highlight further similarities in the (self-)presentation of male/female institutional identities.

5 Concluding remarks

This paper has discussed some of the main theoretical and methodological issues that corpus linguists have to confront when extracting keywords for further analysis, considering in particular the role of keyness measures. We have examined the importance of lockwords, alongside keywords, and have compared the results obtained with different extraction methodologies from two rather different corpora split by gender. Apart from their intrinsic scientific interest, the relevant choices have a major impact on the selection and ranking of the keywords that are then used to conduct research projects, as shown in the illustrative case studies we have presented.

References

Archer, D. (ed.) 2009. *What's In A Word-List? Investigating Word Frequency and Keyword Extraction.* Farnham: Ashgate.

Baker, P. 2004. "Querying Keywords: Questions of Difference, Frequency, and Sense in Keywords Analysis". *Journal of English Linguistics* 32 (4): 346-359.

Baker, P. 2011. "Times may change, but we will always have money: Diachronic variation in recent British English". Journal of English Linguistics 39 (1): 65-88.

Baron, A., Rayson, P. and Archer, D. 2009. "Word frequency and key word statistics in historical corpus linguistics". In Ahrens, R. and Antor, H. (eds.) *Anglistik: International Journal of English Studies* 20 (1): 41-67.

Bondi, M. and Scott, M. (eds.) 2010. *Keyness in Texts*. Amsterdam: John Benjamins.

Bowker, L. and Pearson, J. 2002. *Working with specialized language: A practical guide to using corpora*. London: Routledge.

Cvrček, V. and Fidler, M. 2013. "Not all keywords are created equal: How can we measure keyness?" Paper presented at the *International Conference Corpus Linguistics 2013*. 23 July 2013, Lancaster University, UK.

Drew, P. and Heritage, J. 1992. *Talk at Work: Interaction in Institutional Settings*. Cambridge, Cambridge University Press.

Gabrielatos, C. and Marchi, A. 2012. "Keyness: Appropriate metrics and practical issues". Paper presented at the *International Conference CADS 2012 – Corpus-assisted Discourse Studies: More Than the Sum of Discourse Analysis and Computing?* 14 September 2012, University of Bologna, Italy.

Gaspari, F. 2013. "The languages of maiden speeches in the British Parliament: first-time speakers at the House of Commons vs. the House of Lords". Paper presented at the international conference *The languages of Politics*. 30-31 May 2013, University of Verona, Italy.

Gaspari, F. 2014. "A phraseological comparison of the official online biographical profiles and award motivations of male vs. female Nobel Prize winners". Paper presented at the international conference *Languaging Diversity*. 9-11 October 2014, University of Catania, Italy.

Kemppanen, H. 2004. "Keywords and Ideology in Translated History Texts: A Corpus-based Analysis". *Across Languages and Cultures* 5 (1): 89-106.

Kilgarriff, A. 2009. "Simple Maths for Keywords". In Mahlberg, M., González-Díaz, V. and Smith, C. (eds.) *Proceedings of the Corpus Linguistics Conference CL2009*. 20-23 July 2009, University of Liverpool, UK. Available online at http://ucrel.lancs.ac.uk/publications/cl2009/171_FullPaper.doc

Scott, M. 2009. "In Search of a Bad Reference Corpus". In D. Archer (ed.) *What's in a word-list? Investigating word frequency and keyword extraction*. Farnham: Ashgate. 79-92.

Scott, M. and Tribble, C. 2006. *Textual patterns: keyword and corpus analysis in language education*. Amsterdam: John Benjamins.

Taylor, C. 2013. "Searching for similarity using corpus-assisted discourse studies". *Corpora* 8(1): 81-113.

A corpus-driven study of TripAdvisor tourist reviews of the Victoria Falls

Lameck Gonzo
Rhodes University
lameckgonzo@yahoo.co.uk

The advent of mass travel around the world has ushered in significant changes in the way travellers plan their holidays. In that regard, TripAdvisor has become the world's largest review website (Fina 2011). The reviews on tourism websites represent the traveller's voice and can, as pointed out by Fina, be regarded as a vital tourist text-type since they are a reflection of the traveller's needs, values and expectations. In effect, such review websites constitute a promotional genre. The decisive role of this tourism promotional genre has already been foregrounded by a number of researchers, notably Capelli (2006) Pierini (2009) who demonstrates how reviews can enhance prospective tourists' desire and interest to visit a destination. However, negative comments on the TripAdvisor website can also prevent potential tourists from visiting a particular destination. The discourse deployed by tourism promotional genres is laden with ideological categories, which according to Dann (2006) as cited by Jaworska (2013), are designed to persuade potential customers and convert them into actual clients by appealing to their needs and personal motivations for travel. Thus, this paper contributes to research on the discourse of tourist travel reviews by examining the linguistic features that are evident in the reviews of the Victoria Falls posted on the TripAdvisor's official website between January 2012 and December 2014. The Victoria Falls are classified by UNESCO as one of the Seven Natural Wonders of the World. The paper is guided by the following questions: (1) What patterns of linguistic features are dominant in the TripAdvisor tourist reviews of the Victoria Falls? (2) Which ideologies, if any, are reflected by these dominant linguistic features? and (3) What does an APPRAISAL analysis of the reviews reveal about tourists' attitudes, feelings, experiences and expectations towards the Victoria Falls in particular and Zimbabwe in general? To analyse the reviews, the paper adopts a corpus-driven approach by explicitly combining the tools of corpus linguistics and critical discourse analysis (CDA), following Baker et al (2008) and incorporating Martin and White's (2005) APPRAISAL framework.

The selection of the above-mentioned period is motivated by a number of significant historical and political developments in Zimbabwe. Notable developments in this regard include the expiry at the end of 2012 of the inclusive government formed in 2009 between Zanu PF and the two Movement for Democratic Change (MDC) parties .The harmonised elections of 31 July 2013 and the co-hosting of the United Nations General Tourism Assembly in Victoria Falls by Zimbabwe and Zambia in August of the same year are also worth mentioning.

The contents of this paper are part of an on-going and much bigger project on the discourse of tourism in Zimbabwe. The paper is relevant in view of the decline of the tourism industry in the country and subsequent efforts to revive it. A number of factors have so far been linked to the decline in tourist arrivals. The land redistribution exercise, for example, embarked upon by the government of Zimbabwe in 2000 negatively impacted on world life conservancies (Manwa 2007). The period after the take-over of commercial farms witnessed a critical shortage of foreign currency and fuel, subsequently leading to the deterioration of tourism facilities (Global Eye, 2002). A further blow to tourism in Zimbabwe was the withdrawal of several international airlines (ibid). Given this background, it is instructive to try and unravel tourist linguistic choices to determine and explore attitudes, feelings and expectations in response to Zimbabwe's marketing and rebranding efforts of her tourist destinations.

In this paper, corpus linguistics offers a quantitative dimension, thereby addressing the question to do with the linguistic patterns dominant in tourist reviews of the Victoria Falls. CDA is applied to pinpoint the specific features of language with ideological implications. As for APPRAISAL, the Attitude subcategory reveals tourist feelings and attitudinal patterns in respect of the Victoria Falls. Initial analysis of the corpus (approximately 60 000 words) was done using the tools of corpus linguistics and by applying Antconc software (Antony 2012) that retrieves data on how words behave in a text and displays that data in different formats based on different aspects of linguistic inquiry. I was mainly concerned with the application of wordlist, concordance and collocation, guided by Baker (2006) and McEnery and Hardie (2012). I generated a wordlist to obtain the lists of words and tokens displayed in terms of their frequency in the corpus. Driven by the underpinning assumption of my study that linguistic choices reveal ideological and attitudinal patterns, the most frequent linguistic features would help in exposing tourists' attitudes, feelings and expectations about the Victoria Falls. While acknowledging that frequency counts can be useful, I had to substantially add value to them through the inclusion of concordance lines. A concordance analysis was performed to establish both statistically and qualitatively the behaviour of

lexical features in order to understand their use in the tourist reviews. After identifying the significant words frequently used in the corpus, I calculated the significant collocates of the words in order to establish their contextual meaning and to uncover the ideological and attitudinal assumptions which they embody. The contextual richness of CDA, from the perspective of Fairclough and Wodak (1997) was exploited by studying not only the historical, social and political developments in which the reviews were authored, but also taking into account the tourists' home countries. The APPRAISAL framework was applied to identify positive and negative issues from the linguistic features dominant in the corpus. To realise this, I selected samples of tourist reviews from the corpus and coded Attitude items as either being positive (+) or negative (-) and by distinguishing them as either being inscribed (explicit) or being evoked (implicit). Since the contributors to the corpus were drawn from a wide range of first, second and to some extent third language speakers of English, I expected a broad spectrum of lexical items to be evident in their description of the Victoria Falls. To answer questions regarding tourists' attitudinal patterns, attention was paid to what the data said about the attraction itself, the activities at it, as well as the social and political environment.

My major concern was with content words such as nouns, verbs, adjectives, adverbs and intensifiers since I wanted to establish whether they were used positively or negatively in the corpus. Understandably, results show that the noun *Victoria* has the highest number of occurrences (253 times) bearing in mind that it is the focus of attention. Reference to the Victoria Falls collocates significantly with adjectives such as *amazing, spectacular*, *beautiful, breathtaking* and *stunning* which feature in the top ten of the most frequently used in this semantic category. Use of such adjectives implies viewing the Victoria Falls for the first time induced a great sense of surprise, wonder and shock. In other words, what the tourists saw was beyond their expectations. However, some tourists, especially from Canada, expressed disappointment with the Victoria Falls as evidenced by their use of the adjective *disappointing.* For them, the Victoria Falls were no better than the Niagara Falls of Canada, which are also classified by UNESCO as one of the Seven Natural Wonders of the world. As anticipated, some adjectives, for example, *friendly, welcoming,* did not make reference to the tourist attraction itself, but instead, to the people and the overall environment. Where ideology is concerned, most of the tourists, regardless of their home countries, value friendship and hospitality from their hosts. Their comments suggest they enjoyed these values in spite of Zimbabwe's economic and political tribulations.

The tourist reviews were characterised by highly evaluative language extolling positive features of the Victoria Falls and the services rendered. These were amplified through a wide variety of Graduation resources (*e.g. real, authentic, best, a must visit*) that align the reader with the views and attitudes of the tourists. The overwhelming presence of such linguistic features suggests tourists were really happy to experience the authenticity and naturalness of the Victoria Falls, which undoubtedly, most potential travellers to the site would also want to enjoy. Predictably, several tourists from countries such as the UK, USA and Australia, for example, where English is spoken as a first language, demonstrated great mastery of linguistic choices in their description. Surprisingly, the corpus reveals a sparing reference to values of safety, security and tranquillity. Linguistic items belonging to this semantic category include *safe, peaceful* and *quiet.* Although more than 90% of the comments were positive, there were negative comments with reference to the entry charges and hotel bookings which some tourists said were *too exorbitant*, *pricey*, *overpriced* or *expensive.*

Despite the negative publicity of Zimbabwe from both private and international media, the overall evaluation of linguistic devices employed in the tourist reviews projects a positive image of the Victoria Falls and the country in general as a tourist destination, In that regard, I encourage fellow linguists to interrogate other promotional genres in order to add to the understanding of the role of language in the marketing and rebranding of a given tourist destination.

References

Baker, P. et al 2008. A Useful Methodological Synergy? Combing Critical Discourse Analysis and Corpus Linguistics to examine Discourses of Refugees and Asylum seekers in the UK Press in *Discourse and Society:* 19 (3), 273-306.

Baker, P. 2006. *Using Corpora in Discourse Analysis*: London and New York: Continuum.

Capelli, G. 2006. *Sun, Sea, Sex and unspoilt Countryside: How the English Language makes Tourists out of Readers:* Pari, Pari Publishing.

Dann, G. M. S. 1996. *The Language of Tourism: A Sociolinguistic Perspective*: Oxon, CAB International.

Fairclough, N. and Wodak, R. 1997. Critical Discourse Analysis in T. A. van Dijk (Ed) *Discourse as Social Interaction:* London: Sage Publications, pp 258-284.

Fina, M. E. 2011. What a TripAdvisor Corpus can tell us about Culture: *The Journal of Intercultural Mediation and Communication*, Vl4 pp 59-80.

Global Eye 2002. Tourism under Threat in Zimbabwe http://www.globaleye.org.uk

Jaworska, S. 2013. The Quest for the 'local and authentic': Corpus-based Explorations into the discursive Constructions of Tourist Destinations in British and German Commercial Travel Advertising. In Hohmann, D, (ed) *Tourismukommunication. Im Spannungsfeld von Sprach-und Kulturkontakt* (Series Arbeiten zur Sprachanalyse) Peter Lang Frankfurt am Main pp 75-100.

Manwa, H. A. 2007. Is Zimbabwe ready to venture into Cultural Tourism*? Perspective on Tourism in Africa* 24(3), 365-47.

Martin, J. and White, P. R. R. 2005. *Language of Evaluation: Appraisal in English: Basingstoke: Palgrave* Macmillan

McEnery, T. and Hardie, A. 2012. *Corpus Linguistics: Method, Theory and Practice:* Cambridge: Cambridge University Press.

Pierini, P. 2009. Adjectives in Tourism English on the Web: A Corpus-based Study Circuloide Linguists, Aplicada ala Communication, 40, 93-116.

Methods of characterizing discontinuous lexical frames: Quantitative measurements of predictability and variability

Bethany Gray
Iowa State University
begray@iastate.edu

Douglas Biber
Northern Arizona University
douglas.biber@nau.edu

Joe Geluso
Iowa State University
jgeluso@iastate.edu

1 Introduction

Much phraseological research has focused on recurrent combinations of two or more words, such as lexical bundles or n-grams. Recently, renewed attention has been devoted to recurrent *discontinuous* sequences, or multi-word units consisting of a 'frame' surrounding a variable slot (e.g., *in the * of, the * of the, as a * of, will be * in, to be * to*). While we refer to these types of recurrent combinations as 'lexical frames', they have also been investigated under the terms of 'collocational frameworks' (Renouf and Sinclair 1991; Butler 1998; Marco 2000), 'phrase-frames' or 'p-frames' (Römer 2010; Fletcher 2003/2004/2011; Stubbs 2007), and simply 'frames' (Eeg-Olofsson and Altenberg 1994; Biber 2009; Gray and Biber 2013). Much of this research has focused on two major issues: how to identify recurrent sequences with variable slots, and how to measure the strength of the association between the words that make up the frame.

A third area of inquiry in frame research is the relationship between the frame itself and its possible 'fillers' (the words that occur in the variable slot), evaluating the extent to which a frame is variable or fixed. Because frames are by definition sequences with a variable slot, frame research has been concerned with characterizing discontinuous patterns according to how variable they are (the number of different words that occur in the variable slot), and how predictable the variable slot is (the frequency of the most frequent filler). Previous research has relied primarily upon type-token ratios, but other statistical measures (e.g., entropy, mutual information, proportions) have also been proposed to measure predictability and variability. Yet little research has systematically compared these measures to evaluate the relationships between the measures, or how the properties of the frames themselves may impact what the different measures

are able to capture. The present study attempts to systematically investigate these issues.

2 Methods

We base our analysis on large corpora of conversation and academic writing (c. 4.5 and 5.3 million words respectively) from the Longman Corpus of Spoken and Written English (Biber et al. 1999). Using specialized computer programs, we calculate type-token ratios, proportion of the frame occurrences with the most frequent filler, proportion of frame occurrences with unique fillers (i.e., a filler which occurs only once), mutual information, entropy, and Δp (delta p, Gries 2013) for c. 550 4-word frames (patterns 1*34 and 12*4) that occur at least 40 times per million words in these corpora (identified in Gray and Biber 2013).

3 Results and Discussion

We directly compare results based on these different measures, interpreting what each indicates about types of multi-word associations in linguistic terms, and considering the correlations between the various measures. For example, the results for 12*4 frames in academic writing show that the least predictable frames, based on the proportion of frame occurrences accounted for by the most frequent filler) include patterns with conjunctions within the frame (*between the * and, both the * and*), as well as very frequent preposition-based frames whose variable slot is typically filled by nouns (e.g., *to the * of, as the * of*). Furthermore, a strong negative correlation is found between that measure (i.e., the proportion of the most frequent filler) and type-token ratios. We discuss the implications of findings such as these for the range of measures investigated, demonstrating that different measures have the potential for revealing different types of frames, and different types of associations between frames and their fillers.

We then apply these measures in a comparison of frames across registers, to investigate the differing nature of formulaic language in conversation and academic writing. The results show that frames in the two registers exhibit wide ranges of variability and predictability, with both highly variable and highly predictable frames attested in both conversation and academic writing. However, at the same time, a consideration of the central tendencies for all frames investigated in this study reveals clear differences in the typical frames in the two registers. Table 1 displays selected measures that illustrate these trends for four of the measures (proportion of frame accounted for by the most frequent filler, Δp values for the frame cueing the filler, the type-token ratio, and the proportion of the frame occurrences with a unique filler):

Register	*Pattern*	% most frequent filler	ΔP (filler | frame)	type-token ratio	% of frame with unique filler
ACAD	1*34	19%	0.11	0.41	30%
	12*4	14%	0.10	0.43	31%
CONV	1*34	46%	0.29	0.20	14%
	12*4	44%	0.29	0.19	12%

Table 1. Mean values all frames in academic writing and conversation for selected measures

Table 1 demonstrates that frames in academic writing are typically less formulaic: they are more variable (with higher type-token ratios and proportion of frames with unique fillers) and less predictable (with lower proportions accounted for by the most frequent filler and lower Δp values) than frames in conversation. In contrast, frames in conversation tend to be more formulaic, with higher proportions of frames accounted for by the most frequent filler, stronger associations between the frame and the filler, lower type-token ratios, and a lower proportion of the frames occurring with unique fillers. Thus, this study confirms many of the major register patterns regarding the nature of formulaic language observed in research on continuous phraseological patterns, this time for discontinuous lexical frames. At the same time, Table 1 shows that these measures reveal this same pattern, suggesting that at least these four measures capture the same underlying characteristic of the frames.

References

Biber, D. 2009. "A corpus-driven approach to formulaic language in English". *International Journal of Corpus Linguistics* 14 (3): 275-311.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. 1999. *Longman grammar of spoken and written English.* London: Longman.

Butler, C. 1998. "Collocational frameworks in Spanish". *International Journal of Corpus Linguistics* 3 (1): 1-32.

Eeg-Olofsson, M., and Altenberg, B. 1994. "Discontinuous recurrent word combinations in the London-Lund Corpus". In U. Fries, G. Tottie, and P. Schneider (eds.) *Creating and using English language corpora. Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora, Zürich 1993*. Amsterdam: Rodopi.

Fletcher, W. 2003/2004/2011: online. Phrases in English. Available at: http://phrasesinenglish.org/ (accessed January 2015).

Gray, B. and Biber, D. 2013. "Lexical frames in academic

prose and conversation". *International Journal of Corpus Linguistics* 18 (1): 109-135.

Gries, S.T. 2013. "50-something years of work on collocations: What is or should be next". *International Journal of Corpus Linguistics* 18 (1): 137-165.

Marco, M. 2000. "Collocational frameworks in medical research papers: A genre-based study". *English for Specific Purposes* 19 (1): 63 – 86.

Renouf, A., and Sinclair, J. M. 1991. "Collocational frameworks in English". In K. Aijmer and B. Altenberg (eds.) *English corpus linguistics.* London: Longman.

Römer, U. 2010. "Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews". *English Text Construction* 3 (1): 95-119.

Stubbs, M. 2007. "An example of frequent English phraseology: Distributions, structures and function". In R. Facchinetti (ed.) *Corpus linguistics 25 years on.* Amsterdam: Rodopi.

That-complementation in learner and native speaker corpus data: modeling linguistic, psycholinguistic, and individual variation

Stefan Th. Gries
University of California, Santa Barbara
stgries@gmail.com

Nicholas A. Lester
University of California, Santa Barbara
nicholas.a.lester@gmail.com

Stefanie Wulff
University of Florida
swulff@ufl.edu

1 Introduction

This paper examines the variable realization of the complementizer *that* in English object-, subject- , and adjectival complement constructions as in (1)-(3).

(1) Nick thought (that) Stefan likes tea.
(2) The problem is (that) Stefan doesn't like tea.
(3) I'm glad (that) Nick likes tea.

While native speakers' choices have been researched intensively (see Torres Cacoullos & Walker 2009 and Jaeger 2010 for recent examples), comparatively little is known about what drives L2 learners' decision to realize or omit the complementizer. The present study seeks to address this gap by elaborating on a recent corpus-based study by Wulff (under review).

2 Recent corpus-based work on *that*-complementation in L2 English

Wulff (under review) presents a contrastive corpus-based analysis of *that*-variation in native English speakers and German and Spanish L2 English. She retrieved 9445 instances from native and intermediate-advanced level learner English, including spoken and written corpora (the *International Corpus of English*; the *International Corpus of Learner English*; and the *Louvain International Database of Spoken English Interlanguage*). All instances were annotated for 12 predictors, including the speakers' L1 background; mode; complement type; structural complexity; clause juncture; and the associative bias of the matrix clause verb as either *that*-favoring or zero-favoring (as expressed in a Delta *P* association measure).

In the present study, we revisit Wulff's data with an eye to improving the analysis. More specifically,

there are three different areas in which we go beyond previous work, each of which we briefly discuss below.

3 Extension 1: Surprisal

Following recent research on alternations in speech production in general and *that*-complementation in L1 English in particular (e.g., Jaeger 2010) suggests that one relevant predictor of linguistic choices is the degree to which upcoming linguistic material is (un)expected, or surprising. A frequent operationalization of the notion of surprisal and its effect on processing is, therefore, the negative log of a conditional probability such as $-\log_2 p$ (later material | earlier material). We computed surprisal values for all transitions between words that have or have not been interrupted by the complementizer (and separately so for spoken and written data from the entire BNC) to consider in more detail to what degree predictability of upcoming material affects complementizer realization and the degree to which such affects differ across differently proficient L2 populations. In particular, we are testing the hypothesis that *that* is inserted in cases when the word following *that* is not particularly expected given the word preceding *that*. In additional follow-up work, we are also testing corpus-linguistic association measures from the associative learning literature (Delta *P*) and Kullback-Leibler divergences.

4 Extension 2: Individual Variation

In language acquisition research generally speaking, the considerable impact that individual variation may have on language development has long been recognized. However, with very few exceptions, even the more recent regression-based learner corpus research does not take speaker-specific idiosyncrasies into consideration. In our study, we have now added individual speaker codes to the original data to be able to pinpoint the potential distortions that individual speakers may contribute to the overall regression results. In addition to using a regression approach newly-developed for the corpus-based study of learner language or varieties (see below), we will also be among the first to use mixed-effects modelling approaches in learner corpus research.

5 Extension 3: MuPDAR

The final extension has to do with how one can best target the difference between native and non-native speaker behavior. Recent studies in learner corpus research (Gries & Deshors 2014, Gries & Adelman 2014, Wulff & Gries to appear) have developed an approach called MuPDAR (for Multifactorial Prediction and Deviation Analysis with Regressions) that is specifically designed to target to what extent, in what way(s), and why non-native speakers make linguistic choices that align with those of native speakers in comparable speech situations. This procedure involves the following steps:

- a first regression in which one models native speaker choices only;
- the application of that regression to the non-native speaker data to see how the actual non-native speaker choices compare to the predicted nativelike choices;
- a second regression in which one determines which predictors of an alternation give rise to non-nativelike choices.

It is this approach that we will apply to the otherwise already annotated data (but see below).

6 Initial results

The results of a first regular binary logistic regression analysis are promising: The minimal adequate regression model (LR=5059.45, df=28, p=0) predicts all speakers' choices very well (Nagelkerke's R^2=-.55, C=0.88; classification accuracy=80.63%). The results suggest that (i) processing-related factors most strongly impact native speakers' and learners' choices alike; (ii) Spanish learners are more conservative regarding complementizer omission than German learners; (iii) both learner groups exhibit knowledge of target-like verb-complement type associations; and (iv) both learner groups indeed display sensitivity to register differences.

We are currently in the process of running the second analysis with the three extensions, which make this the first learner corpus study that features the notion of surprisal as well as the combination of mixed-effects modeling and MuPDAR. Preliminary results of this more refined analysis indicate that, as is not uncommon, the mixed-effects modeling approach increases the classification accuracy considerably (given how individual-speaker variation is accounted for), but that the effect of surprisal is less strong than would have been expected on the basis of corresponding native speaker data (e.g., Jaeger 2010).

References

Gries, St.Th. and Adelman, A.S. 2014. "Subject realization in Japanese conversation by native and non-native speakers: exemplifying a new paradigm for learner corpus research". *Yearbook of Corpus Linguistics and Pragmatics 2014: New empirical and theoretical paradigms*. Cham: Springer.

Gries, St.Th. and Deshors, S.C. 2014. "Using regressions

to explore deviations between corpus data and a standard/target: two suggestions". *Corpora* 9 (1): 109-136.

Jaeger, T.F. 2010. "Redundancy and reduction: Speakers manage syntactic information density". *Cognitive Psychology* 61: 23-62.

Torres Cacoullos, R. and Walker, J.A. 2009. "On the persistence of grammar in discourse formulas: a variationist study of *that*". *Linguistics* 47 (1): 1-43.

Wulff, S. and Gries, St.Th. to appear. "Prenominal adjective order preferences in Chinese and German L2 English: a multifactorial corpus study". *Linguistic Approaches to Bilingualism*.

Wulff, S. under review. "A friendly conspiracy of input, L1, and processing demands: *that*-variation in German and Spanish learner language". Submitted to: A. Tyler, L. Ortega and M. Uno (eds.) *The usage-based study of language learning and multilingualism* (Proceedings of GURT 2014). Georgetown: Georgetown University Press.

Recent changes in word formation strategies in American social media

Jack Grieve
Aston University
j.grieve1@aston.ac.uk

Andrea Nini
Aston University
a.nini1@aston.ac.uk

Diansheng Guo
University of South Carolina
guod@mailbox.sc.edu

Alice Kasakoff
University of South Carolina
kasakoff@mailbox.sc.edu

1 Introduction

Current linguistic research is focusing more and more on social media such as Facebook or Twitter. Even though many studies on language variation and change have been carried out using large corpora of social media texts (e.g. Eisenstein et al. 2012; Doyle 2014), not many studies have focused on the analysis of new words emerging from these social media and on their characteristics. This paper begins to fill this gap by presenting results of a study on the emerging new trends of word formation in a large corpus of Twitter messages produced in America.

2 The corpus

The corpus used for this study consists in 6 billion word tokens of geo-coded American tweets collected between January and September 2013 using the Twitter API. The collection of tweets involved only those tweets that were produced within the contiguous United States and that contained both a timestamp and a geocode with the longitude and latitude for the location where the tweet was sent.

3 Methods

As a first step, the 60,000 word types that occurred in the corpus at least 1,000 times were extracted. Through this step it was possible to remove those word types that occurred rarely in the corpus. After this step, the relative frequency of occurrence of the remaining types was calculated for each day represented in the corpus. A Spearman rank-order correlation coefficient was then calculated between the relative frequency of each of these 60,000 word types and the day of the year. By ordering the word types by the value of Spearman rho it was possible to observe which word types increased and which word types decreased in frequency in American tweets during 2013.

4 Results

As an example of the results, the 10 strongest positive and negative Spearman correlation coefficients are presented in Table 1.

Increasing	Decreasing
rn (.978)	*wat* (-.976)
selfie(s) (.965)	*nf* (-.962)
tbh (.960)	*swerve* (-.956)
fdb (.952)	*shrugs* (-.956)
literally (.948)	*dnt* (-.956)
bc (.943)	*wen* (-.948)
ily (.940)	*rite* (-.947)
bae (.934)	*yu* (-.946)
schleep (.932)	*wats* (-.946)
sweg (.932)	*yeahh* (-.945)

Table 1: Top increasing and decreasing words

Among the increasing words, we observe new coinages, such as *selfie(s)* (a photo of oneself), *schleep* (sleep), *sweg* (swag) and *bae* (babe). The word *literally* was also found on the increase and among the several likely explanations for this increase it is possible to propose the rise of a new meaning of *literally* or the increase in the formality of tweets over time. In general, however, acronyms were on the rise in the corpus. Examples of acronyms are *rn* (right now), *tbh* (to be honest), *fdb* (fuck dem bitches) and *ily* (I love you).

Among the decreasing words we observe a number of creative spellings of already established word types, such as *wat* (what), *nf* (now following), *dnt* (don't), *wen* (when), *rite* (right), *yu* (you), and *wats* (what's).

The distributions of these frequencies over time were also explored through scatterplots. As an example of the general patterns observed, the distribution of the top two increasing words, *rn* and *selfies*, and of the top two decreasing words, *wat* and *nf*, are reproduced in, respectively, Figure 1, Figure 2, Figure 3 and Figure 4.

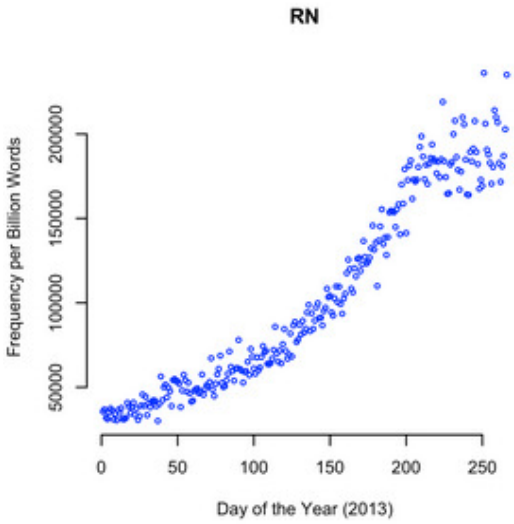


Figure 1: Relative frequency of *rn* over time

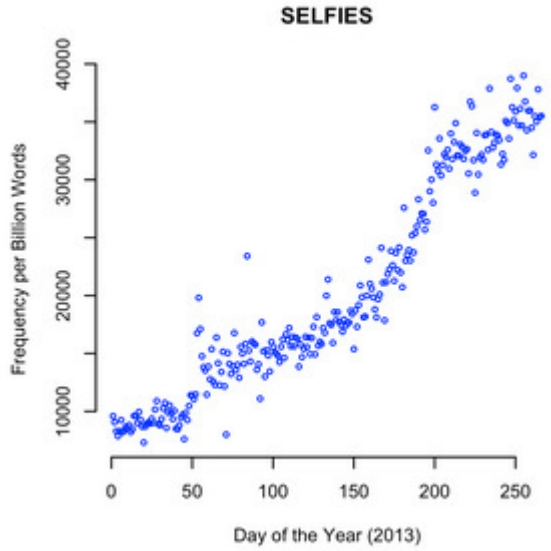


Figure 2: Relative frequency of *selfies* over time

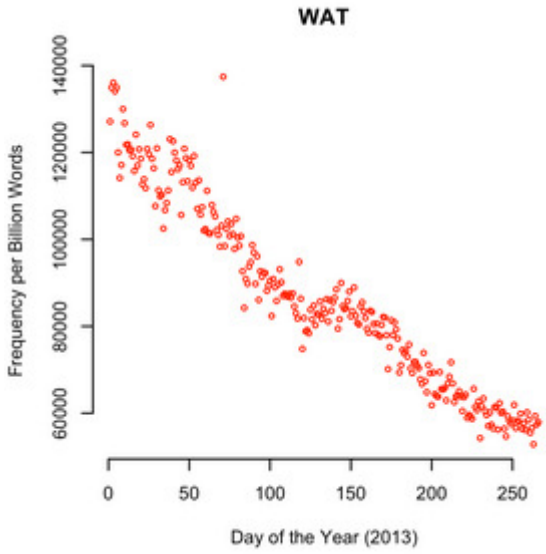


Figure 3: Relative frequency of *wat* over time

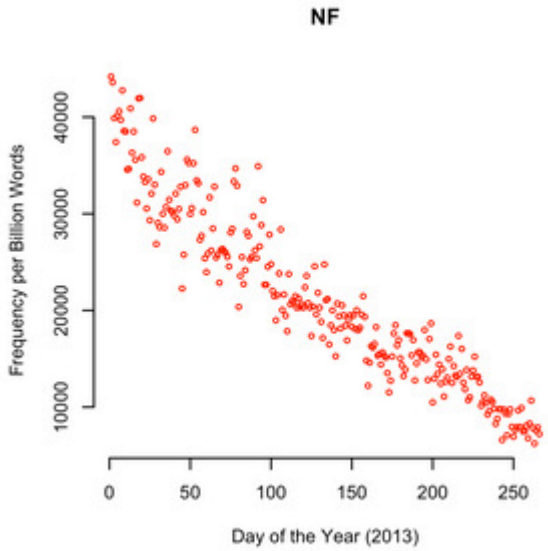


Figure 4 : Relative frequency of *nf* over time

Visual inspection of the examples above and of the other scatterplots suggests that the emerging of new words follows an s-shaped curve of diffusion (Rogers, 2003) whereas the decline of old words follows a steadier and almost linear decrease.

5 Discussion

The results of the study show that the use of creative spellings decreased in American tweets produced during the year 2013, while the use of acronyms increased. The constraints of social media in general are likely to push their users to produce shorter forms and, consequently, to innovate language by introducing new forms that are able to express meaning in few characters. It is clear that both creative spellings (e.g. *wat*, *rite*, *wen*) and acronyms (*rn*, *tbh*, *ily*) achieve this purpose. However, the analysis suggests that acronyms are substantially more popular strategy in contemporary micro-blogging, presumably the 140 character constrain imposed by Twitter has forced the users to contract meaning even more by using acronyms that can reduce common multi-word fixed or semi-fixed phrases in few characters. If this hypothesis is correct, the two trends reflect a change of habit that developed through time and that was determined by the medium.

Apart from the findings on the type of word formation, the examination of the scatterplots suggests that these changes in word formation patterns mirror other types of linguistic change, such as phonological and syntactic changes by following an s-shaped curve typically found in sociolinguistics (Labov, 1995). So far, however, the s-shaped curve has been found in cases of alternation variables in which either the presence or absence of a change is recorded. The present study has produced initial findings that suggest that the same s-shaped curve can be found when frequency variables are considered, with the limit of growth of the curve consisting in the upper limit of that meaning being discussed within the speech community under analysis. A similar s-shaped curve is often found in the diffusion of new symbols or behaviours in various other spheres, such as technology, news, fashion and other aspects of cultural phenomena that represent innovations (Rogers, 2003). The exploration of the similarity between the mechanisms of the diffusion of innovative linguistic items and the diffusion of innovations in other aspects of society is important to be pursued in the future.

6 Conclusions

The present paper reports on an analysis of six billion of tweets for change in patterns of word formation. The results of the study are two-fold. Firstly, it was found that in 2013 American tweets started a shift from a creative spelling word formation trend to an acronym word formation trend. The 140 character medium constrain of Twitter as well as the increasing degree of information used in Twitter has been proposed as main explanation of this phenomenon. Secondly, the words on the increase show an s-shaped pattern that is typical of linguistic changes. If replicated in future studies, these findings can have significant implications for the understanding of the effect of the medium on language evolution and change.

Acknowledgements

This research is funded by the Economic and Social Research Council, the Arts and Humanities Research Counties, and JISC in the United Kingdom and by the Institute of Museum and Library Services in the United States, as part of the *Digging into Data Challenge*.

References

Doyle, G. (2014) Mapping dialectal variation by querying social media, In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics.

Eisenstein, J., O'Connor, B., Smith, N. and Xing, E. (2012) Mapping the geographical diffusion of new words, arXiv:1210.5268 [cs.CL], pp. 1–13, Available from: http://arxiv.org/abs/1210.5268 (Accessed 13 June 2014).

Labov, W. (1995) Principles of Linguistic Change. Volume I: Internal Factors, Oxford, Blackwell.

Rogers, E. M. (2003) Diffusion of Innovations, New York, Free Press.

Smith, A. and Brenner, J. (2012) Twitter use 2012, Pew

Internet & American Life Project, Available from: http://www.looooker.com/wp-content/uploads/2013/05/PIP_Twitter_Use_2012.pdf (Accessed 21 July 2014).

Transgender identities in the UK mainstream media in a post-Leveson context

Kat Gupta
University of Nottingham
kat@mixosaurus.co.uk

1 Introduction

In this paper I examine the media representation of two trans women, Lucy Meadows and Chelsea Manning. Both women were widely reported in the UK mainstream press in 2013, a period coinciding with Part 1 of the Leveson Inquiry. This Inquiry examined the culture, practices and ethics of the press, with a particular focus on the relationship between the press and the public, the police and politicians.

The Inquiry heard evidence from "organisations representing minority, community and societal groups alleging that individuals within those groups, or the groups themselves, have attracted inaccurate and discriminatory press interest" (Leveson 2012: 448); among these were two written and one oral submission from the organisation Trans Media Watch. In these, they described patterns of negative media representation of trans people, including routine use of previous names, routine use of "before" photos, demeaning and intimidating language for comic effect, and misgendering.

I argue that press misgendering can take more subtle forms than the reporter's use of "quotation marks to dismiss the veracity of the subject's identity inappropriate pronouns or placing the person's identity in" (Trans Media Watch 2011: 11). I examine press usage of pronouns in direct quotations and repetition to investigate how these can be used to undermine trans people's identities.

2 Context

The term "transgender" or "trans" is used as an umbrella term to cover a wide range of gender identities including those of trans women, trans men and people with non-binary, genderfluid and agender identities. Trans people usually experience a sense of misalignment with the sex they were assigned at birth and the gender they identify as; this is in contrast to "cisgender" or "cis" people whose assigned sex and gender identity are aligned. I focus on the experiences of trans women – people who were assigned a male sex at birth, but who identify and/or live as women (Serrano 2007: 11). Serrano (2007: 12) argues that trans women face a complex

interaction of transphobia, cissexism[47] and misogyny which create a culture where trans women are hyperfemininised, trans women are hypersexualised and sex reassignment surgery is sensationalised.

Serrano argues that the media is a crucial component in creating a culture in which trans women's identities are routinely dismissed. This is supported by research by Trans Media Watch (2011) which was submitted and subsequently included in the Leveson Inquiry report (2012: 448):

> transgender people are subject to disproportionate and damaging press attention simply by dint of being members of that group, rather than in consequence of anything they might have said or done, and because of what they describe as an obsession in parts of the British press with 'outing' members of the transgender community

It is important to note that negative media representation has a devastating effect on an already vulnerable population. The Trans Mental Health Study (McNeil et al. 2012) found that 92% of respondents had heard that trans people were not normal (McNeil et al. 2012: 41) and that 84% of respondents had thought about ending their lives at some point (McNeil et al. 2012: 59). Trans Media Watch (2011: 8) highlighted the effect of negative media representation on respondents:

- 67% of respondents said that seeing negative items in the media about transgender people made them feel "angry".
- 51% said that these items made them feel "unhappy".
- 35% said that they felt "excluded".
- 20% said that they felt "frightened

As these figures indicate, negative media portrayals of trans people have consequences.

3 Case study: Lucy Meadows

In March 2013, a woman named Lucy Meadows was found dead at her home. Meadows, a primary school teacher, was transitioning from male to female. In December 2012, the school announced her decision to return to work after the Christmas break as Miss Meadows. This was reported in the local press and quickly picked up by the national press. Three months later, Meadows was found dead. Her death prompted discussions of press freedom, the contributions of trans people to society and responsible media representation of trans lives and experiences.

I use two corpora: a small, focused corpus (166 texts, 108,643 words) of news texts reporting on Lucy Meadows between October 2012 and October 2013, and a reference corpus (7000 texts, 3,954,808 words) of news texts sampled from the same time period. The gendered pronouns *she* and *her* emerged as key terms. By examining gendered pronouns when used to refer to Meadows, I found that *he* was overwhelmingly used before Meadows' death – when she had already expressed her intention to live and work full-time as female. Media reporting of Meadows' transition appears to dismiss her gender identity in favour of presenting her as the sex she was assigned at birth. This finding appears to reinforce observations by Serrano and Trans Media Watch.

	he	*she*
Before death	124	20
After death	38	451

Table 1: Pronoun use before and after Meadows' death

However, as Table 1 shows, female pronouns were overwhelmingly used after her death. This is probably due to several factors, not least campaigns for improved reporting on trans issues by activists.

The data indicates that, while tabloid misgendering is an issue, the situation is complicated by use of direct and indirect quotations. Direct quotations account for 65 of the 124 occurrences of *he* before death and 10 of the 38 occurrences of *he* after death. Of these, repetition accounted for a considerable percentage of occurrences – there were 33 occurrences of a single sentence ("he's not only in the wrong body...he's in the wrong job") from an article by Richard Littlejohn, a columnist from the *Daily Mail*. However, not all of these repetitions were uncritical reproductions of Littlejohn's writing. Instead, journalists were criticising Littlejohn but in doing so, were also reproducing transphobic text.

4 Case study: Chelsea Manning

Chelsea Manning announced her female gender identity in a press release issued on 22 August 2013, the day after she was sentenced for leaking classified material to WikiLeaks. Manning remains a polarised figure. As a high profile prisoner accused of aiding the enemy, espionage and theft, she attracted fury – yet, for exposing American abuses of power, she was viewed as a hero by others and was nominated for the Nobel Peace Prize in 2014.

The Human Rights Campaign supported her, arguing that her "transition deserves to be treated

[47] Defined as a "belief that transsexuals' identified genders are inferior to, or less authentic than, those of cissexuals" (Serrano 2007: 12). This often manifests as denying trans people the treatment associated with their identified gender; Serrano offers the examples of using the wrong pronouns or forcing the trans person to use different toilets.

with dignity and respect and that "[u]sing the name Bradley or male pronouns is nothing short of an insult" (Krehely 2013). However, Manning's announcement meant that she was reported as a trans figure and therefore subject to the negative media representation discussed earlier. Trans Media Watch's (2013) response highlighted BBC reports as particularly concerning.

In this stage of research, I use a small, focused corpus of news texts reporting on Chelsea Manning between August 2013 and October 2013. Crucially, Manning's transition was reported eight months after Meadows', and in an environment with increased awareness and support for good practice in reporting trans issues. By examining gendered pronouns when used in reference to Manning, I am able to further explore these more subtle manifestions of mispronouning.

5 Conclusions

While corpus linguistics have been used to explore gender (c.f. Baker 2008, 2014), these have tended to focus on (self-)construction of cisgender identities[48] with little attention given to transgender identities. By examining transgender identities as constructed by the mainstream UK press, I am able to investigate issues of minority representation, press tactics of negative representation, and the interactions between press, public, reporters and reported. I demonstrate that mispronouning, while a key part of negative media portrayal used to dismiss trans peoples' gender identities, is more complex than the hostile use of quotemarks identified by Trans Media Watch. Through repetition of selected direct quotes, the press is able to reinforce some voices and not others. In doing so, reporters are able to evade direct responsibility for misgendering while continuing to produce the effect of undermining a trans person's gender identity.

References

Baker, P., (2005). *Public Discourses of Gay Men*. London: Routledge

Baker, P. (2008). *Sexed Texts: Language, gender and sexuality*. London: Continuum.

Baker, P. (2014). *Using Corpora to Analyze Gender*. London: Bloomsbury.

Krehely, J. (2013). Pvt. Chelsea E. Manning Comes Out, Deserves Respectful Treatment by Media and Officials. Retrieved from http://www.hrc.org/blog/entry/pvt.-chelsea-e.-manning-comes-out-deserves-respectful-treatment-by-media-an

Leveson, B. (2012). *An inquiry into the culture, practices and ethics of the press*. London: The Stationery Office.

McNeil, J., Bailey, L., Ellis, S., Morton, J. and Regan, M. (2012). The Trans Mental Health Study. Retrieved from http://www.gires.org.uk/assets/Medpro-Assets/trans_mh_study.pdf

Serrano, J. (2007). *Whipping Girl: a transsexual woman on sexism and the scapegoating of femininity*. Berkley: Seal Press

Trans Media Watch. (2011). *The British Press and the*

Transgender Community: Submission to The Leveson Inquiry into the culture, practice and ethics of the press. Retrieved from http:/www.levesoninquiry.org.uk/wp-content/uploads/2012/02/Submission-by-Trans-Media-Watch.pdf

Trans Media Watch. (2013). *Trans Media Watch responds to Chelsea Manning coming out*. Retrieved from ttp://www.transmediawatch.org/Documents/Press_Release-20130822.pdf

[48] With exceptions such as Baker's (2005) examinations of House of Lords reform on the age of consent and British tabloid representation of gay men.

Lexical selection in the Zooniverse

Glenn Hadikin
University of Portsmouth
Glenn.hadikin@port.ac.uk

1 Introduction

In this paper I will attempt to formalise the study of language and memes with a hypothesis I call Lexical Selection (Hadikin 2014). Closely based on Dawkins (1976) concept of memes and Hoey's (2005) Theory of Lexical Priming I discuss the idea that strings of lexical items enter into a competitive environment in which they compete with rival strings in the discourse community.

The theoretical concept will be applied to data from an online science forum - Zooniverse - part of a community of 'citizen scientists' that work with scientific images in their free-time to classify galaxies, hunt near-earth asteroids and transcribe museum records (three of 25 affiliated projects). In this abstract I focus on the lexical environment around the item *I* to explore what it can tell us about the site users.

2 Lexical Selection

The Lexical Selection hypothesis argues that language strings acts as replicators in a Dawkinsian sense (Dawkins 1976). Figure 1 shows a simple representation of the process.

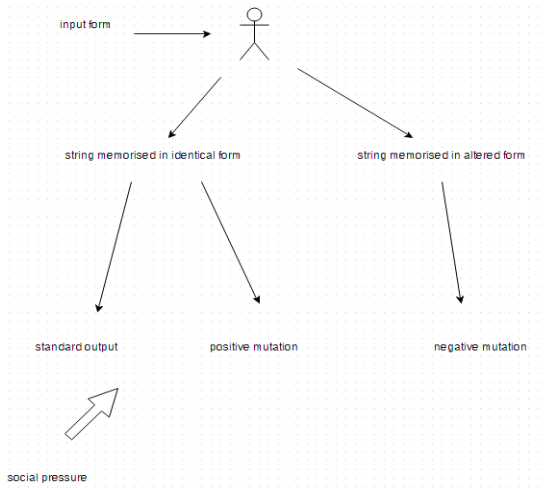


Figure 1: Lexical selection process reproduced from Hadikin (2014)

Consider a string 'the accuracy of the parser' taken from a paper in the ACL corpus - an archive of computer linguistics publications available via SketchEngine (Kilgariff *et al*. 2014). A proficient writer (or, indeed, a proficient speaker) is likely to be primed to use this structure to refer to the accuracy of a parser that is known to the audience. They have the string stored mentally in 'identical form' (technically a mental representation of every previous occurrence the writer has encountered). In future situations the writer will be free to select from 'standard output' - i.e. to reuse the same form or to create a positive mutation. This may well take the form 'the accuracy of our parser' as an example. Alison Wray (personal communication) noted that this is certainly not a random process but, nonetheless, it generates an alternative form that the next 'generation' of readers and listeners (i.e. anyone who reads or listens to the writer) may choose.

The right side of figure 1 represents a situation where a language learner does not appear to store the identical form. In this case certain parts of the structure may be lost. One example in the ACL corpus is 'the accuracy of Japanese parser'. I have labelled these forms 'negative mutations' because there is a strong chance that social pressure - supervisors, reviewers, editors and the like - will remove the form from the lexical gene pool but, as this string has survived to be in the corpus, it may well be surviving and could lead to replication in certain communities.

3 Zooniverse data

Zooniverse is the world's largest 'citizen science' or science crowdsourcing website with over 1.2 million registered users at the time of writing. Site members engage in simple tasks such as looking at an image of a galaxy from a bank of images and matching it with a preset range of shapes offered by the site owners. The cumulative efforts of the users - or Zooites - leads to real scientific results. One example is the discovery of a rare object called Hani's Voorwerp by a teacher called Hanny van Arkel. These are thought to be the remnants of galaxies and are now called 'Voorwerpjes' in her honour (van Arkel 2014). The data used in this short study is a single discussion thread about dark matter with 204 000 word tokens.

4 Study of *but I*

One aspect of the Zooniverse data I wanted to explore is the selection of pronouns and any potential units that may carry them. To begin this I begin with the pronoun *I*. The most frequent *I* patterns shown in the data (using WordSmith tools 6, Scott 2012) is shown in figure 2.

but		have
and	I	'm
as		am

Figure 2: Most frequent L1 and R1 items co-occuring with *I* in dark matter corpus

The rest of this abstract will briefly sketch my analysis of the string *but I* - if and how it might be seen as part of a lexical selection system.

There are a total of 80 concordance lines for *but I*; I started reading each one for more detail about what was being discussed and what concepts were being contrasted (if, indeed, that was the case at all).

Consider a single line of data as an example -

```
I wholly agree [Zooite's name] but I rest easy and smile. It may be a single phenomenon but it's a huge one
```

Here the writer expresses agreement with a previous poster's call for caution in a calculation. With this moderate amount of data I have been able to go back and look at the context in detail. This is important because the poster will know the context well and, in line with Hoey's Lexical Priming (2005), will be primed to make selections accordingly.

self	44
general statement	25
other Zooites	7
other people or organisations	2
we	1
other functions of 'you'	1

Figure 3: Categories of reference directly preceding *but I* in dark matter corpus

As highlighted in figure 3 I placed the 80 lines of data into six categories depending on what was being referred to directly before *but I* appears. A notable difference between the *self* set and the *general* set is five lines of *but I do* data as follows -

```
them out, but I do it deliberatel
 is false but I do believe that s
all size but I do come up with di
 ratios but I do operate at a hig
o ... but I do prefer to organise
```

Note that these five posts were all from a single user. We appear then, to be looking at this user's idiosyncratic primings and must be duly cautious about any claims. The writer - I will give him a pseudonym Trevor - is using two kinds of structure here. The first employs *do* as a main verb in the expression *but I do it deliberately* and the second uses emphatic *do* - an auxiliary verb that adds emphasis to the point being made such as *but I do believe that.* It would take a different kind of data to investigate whether Trevor uses these exact strings (or related frames) in other situations and domains. The 28 lines of *DO it deliberately* in the BNC suggest the core of a shared unit in everyday UK English; this raises interesting questions about how culturally shared units nest and interact and will need to be explored in further work.

This dataset is also unusual for other reasons. Firstly, there is not a single occurrence of *but I do* in 25 lines where a general statement (usually about physics) precedes the 2-gram. This might suggest a new aspect of Lexical Priming or the need for further discussion of the fine line between semantic and pragmatic association (see Hoey 2005).

Secondly, there are no occurrences where *but I* combines with *don't* in the *self* set. Returning to the full 80 lines of *but I* data we see 8 occurrences of *but I do* compared with a single occurrence of *but I don't*. A 250 line sample of a corpus of online texts (enTenTen 2013 available via Sketch Engine), however, has 22 occurrences of *but I do not*, 151 occurrences of *but I don't* and just 77 of *but I do*. A ratio of just over 2:1 in favour of the negative structures. This suggests that the pragmatic environment and subject matter is influencing the choice of units in interesting ways - clearly in a situation where the user writes -

```
the Vatican goes along with it but I don't suppose they understand the second law
```

(when discussing the big bang) they would not likely consider *but I do* as an option but could reasonably make the point without using a *but I* form at all.

5 Conclusion

To conclude, there is little doubt that units of language are moving and being shared in some form from generation to generation - our task, as researchers, is to explore the nature of such units and to learn more about the mechanism. The Zooniverse provides a wealth of data and the right questions about its users could lead to a better understanding of motivation and why over 1.2 million people give up their free time helping scientists and other research teams. In this short work in progress one influential user has had a notable and complex effect on this data set but the role and, indeed, language choices of such influential individuals may prove to be a significant aspect of the linguistics of social media and crowdsourcing websites. The Lexical Selection hypothesis provides us with an additional perspective on language that has sometimes been neglected in Corpus Linguistics and has the potential to provide new insights as it encourages us to remember the human actors behind our data.

References

Dawkins, R. 1976. *The Selfish Gene*. Oxford: Oxford University Press.

Hadikin, G. 2014. *Lexical Selection and the Evolution of Language Units.* Manuscript submitted for publication.

Hoey, M. 2005. *Lexical Priming: a New Theory of Words and Language*. London: Routledge.

Kilgariff, A. 2014. 'The Sketch Engine: Ten Years on.' *Lexicography 1 (1)*: 7-36.

Scott, M. 2012. *WordSmith tools version 6*. Liverpool: Lexical Analysis Software.

van Arkel, H. 2014. *Voorwerp*. Available online at http://www.hannysvoorwerp.com.

'In typical Germanic fashion': A corpus-informed study of the discursive construction of national identity in business meetings.

Michael Handford
Tokyo University
mjahandford@gmail.com

Corpus methods have been effectively employed in several language-related areas and disciplines, such as discourse analysis (Baker, 2006; Stubbs, 1996), professional discourse (Koester, 2006; Handford, 2010), and translation studies (Baker, M., 1995). One field of study that could further employ a corpus methodology is intercultural communication studies (ICS).

A useful distinction in ICS contrasts 'intercultural', implying some degree of interaction between members of differing groups, with 'cross-cultural', comparing communicative behaviours of different cultural groups. Although corpus tools have received little attention in the former type of study (Handford, 2015), they have been effectively employed in 'cross-cultural' comparisons involving parallel and comparative corpora, for example Biber et al's (1998) comparison of US and UK language usage, Connor et al's (2008) study of Spanish and US pharmaceutical labels, and Fahey's (2005) study of apologies in Chilean and Irish soap operas. An innovative study by Stubbs (1996) explored a large corpus of British English to reveal the way certain 'cultural keywords' (that is words that are deemed to be salient within that particular culture, such as 'democracy') are used across various contexts.

While such studies have led to interesting findings, they are all predicated on a 'received culture' perspective that sees culture as a given, rather than a concept that requires explanation. For instance, Stubbs (1996: 181) uncritically assumes that 'British culture' is a tangible, discoverable object. This reified, commonsense view of culture (e.g. Hofstede, 1991) is arguably essentialist in nature, and contrasts with much discourse-influenced work into sociocultural identity (e.g. Bucholtz and Hall, 2005; Benwell and Stokoe, 2006) and interculturality (Collier and Thomas, 1988; Dervin, 2012; Handford, 2014) which see culture and (socio)cultural identities as constructed, emergent and negotiated in and through discourse. Thus, while corpus-based ICS have largely approached culture as a given, and tend to conflate culture with nationality, there is nothing inherent in a corpus methodology that necessitates such an approach. Indeed, if we accept the complementarity

of corpus linguistics and discourse analysis (Baker, 2006), corpora should lend themselves to analysing the dialogic emergence of culture and identity.

This talk employs a methodology that combines corpus tools with qualitative discourse-analysis methods to analyse interculturality in business meetings (Handford, 2014, 2015). Rather than using the corpus as a repository of examples which the researcher draws on to highlight differences between reified cultural groups, this approach can illuminate the discursive construction of sociocultural identities independent of the researcher's stance and stereotypes. The corpus used is the Cambridge and Nottingham Business English Corpus (CANBEC), a one-million word corpus of primarily authentic business meetings (see Handford, 2010; copyright Cambridge University Press). The meetings are from a wide range of contexts, involving business interactions between speakers of differing professions, organisations, industries, nations and local teams. In other words, they involve speakers from both large and small cultures (Holliday; 1999).

The methodology combines Gee's notion of 'situated meaning' (Gee, 2005), that is the utterance-token indexical meaning of a lexicogrammatical item in a specific context, with discourse prosody, to analyse both statistical keywords (Scott, 2011) and cultural keywords (Stubbs, 1996, following Williams). Situated meanings can index specific sociocultural identities, and by exploring the discourse prosody of repeated invocations of the same identity through the same item or category, the combination of corpus and discourse methods can arguably achieve a synergistic result. The depth of understanding of discourse analysis is combined with the breadth of analysis offered by corpus methods.

Specifically, the top statistical keyword in business meetings, *we*, (Handford, 2010) and cultural keywords and phrases denoting nation and nationality (e.g. *Chinese*, *this nation of ours*, etc.) are explored in concordance lines, with iterative reference to the background context and when necessary extended co-text, to ascertain how they can inform our understanding of intercultural communication in professional contexts. In this way, the talk intends to address Piller's (2011:91) exhortation for future research: 'instead of treating national culture as a given, intercultural business communication will need to see the nation as a discursive construction that social actors draw upon in selective ways' – specifically through the analysis of nationality markers in meetings. However, this talk also addresses Piller's other call (2011:92-3) to move beyond seeing culture as equivalent to nationality, through the analysis of the indexical pronoun *we* (Handford, 2014).

The top CANBEC keyword *we* shows that it can index a wide range of sociocultural identities, such as inclusive inter-organisational, exclusive national or inclusive local identities, and such identities are indexed dynamically and emergently through the discourse. A qualitative analysis of three international, inter-organisational, inter-professional meetings shows that by far the most frequently identity indexed by *we* is the organisational, whereas national identity is only indexed in non-transactional exchanges, such as small talk at the beginning of a meeting (Handford, 2014).

With reference to nationality or nation markers, one of the interesting factors is their relative infrequency in the corpus when compared to organisational identity markers. When explicitly indexed, the speaker's nationality, an interlocutor's nationality, nationality as a particular market (e.g. 'we don't really wanna get involved with the Chinese'), and as a national-professional group (the UK fire service) can be signalled. Various functions are invoked through the use of nation/nationality markers in CANBEC, including reporting (e.g. about sales areas), distinguishing between individuals, making relational asides (Koester, 2004), evaluating, justifying, and Othering.

This talk shows that the discourse prosody of *we* is largely positive or neutral, whereas nationality is either negative or neutral. This interpretation of *we* includes the use of exclusive identities, such as exclusive inter-organisational *we* denoting the speaker's organisation but not the other company represented at the interaction, and *we* is arguably one of the ways speakers create the sense of comity and cooperation that is central to effective business. In contrast to *we*, nationality markers are usually negative or neutral, for instance in self-references to nationality ('it's like typical English disease', uttered by an English national), and the Othering examples cited above. Although Othering entails negative discourse prosody, this is not to suggest it causes divergence in the context of use. A similarity between national *we* and explicit nationality markers is their occurrence in non-transactional discourse in meetings, either in small talk exchanges, or in relational asides (Koester, 2004); in other words, nationality can be drawn on as a relational resource to fulfill interpersonal goals in professional contexts.

Such findings have implications for the epistemological status of culture and nationality in ICS and professional communication studies, in that they support the assertion (e.g. Holliday, 1999; Piller, 2011) that we need to see nationality in a more discursive and critical light: as explanandum rather than explanans. The finding that social identities other than the national may be indexed far more frequently adds weight to the argument that

mainstream approaches to intercultural (business) communication may have overemphasized the importance of nationality, although the extent to which corpora and discourse can shed light on such a question is open to debate (Handford, 2014).

References

Baker, M. (1995). Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target*, 7 (2), 223–243.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Benwell, B. & Stokoe, E. (2006). *Discourse and Identity*. Edinburgh: Edinburgh University Press.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure & Use*. Cambridge: Cambridge University Press.

Bucholtz, M., & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, 7 (4-5), 585–614.

Collier, M. J., & Thomas, M. (1988). Cultural Identity: An Interpretive Perspective. In Y. Y. Kim & W. B. Gudykunst (Eds.), *Theories in Intercultural Communication* (pp. 99-120). Newbury Park, CA: Sage.

Connor, U., Ruiz-Garrido, M., Rozycki, W., Goering, E., Kinney, E., & Koehler, J. (2008). Patient-directed medicine labeling: Text differences between the United States and Spain. *Communication & Medicine*, 5 (2), 117-132.

Dervin, F. (2012). Cultural identity, representation and Othering. In J. Jackson (Ed), *Routledge Handbook of Intercultural Communication* (pp. 181-194). Abingdon: Routledge.

Fahey, M. (2005). Speech acts as intercultural danger zones: a cross cultural comparison of the speech act of apologizing in Irish and Chilean soap operas. *Journal of Intercultural Communication*, 8, 1404-1634.

Gee, J.P. (2005). *An Introduction to Discourse Analysis*. Abingdon: Routledge.

Handford, M. (2010). *The Language of Business Meetings*. Cambridge: Cambridge University Press.

Handford, M. (2014). Cultural identities in international, interorganisational meetings: a corpus-informed discourse analysis of indexical 'we', *Language and Intercultural Communication*, (14) 1, 41-58.

Handford, M. (2015). Corpus Linguistics. In Zhu Hua (Ed.) *Research Methods in Intercultural Communication*. Oxford: Wiley-Blackwell.

Hofstede, G. (1991). *Culture and Organisations*. New York: McGraw-Hill.

Holliday, A. (1999). Small Cultures. *Applied Linguistics*, 20 (2), 237-264.

Koester, A. J. (2004). Relational sequences in workplace genres. *Journal of Pragmatics*, 36, 1405–1428.

Koester, A. (2006). *Investigating Workplace Discourse*. Abingdon: Routledge.

Piller, I. (2011). *Intercultural Communication: A Critical Introduction*. Edinburgh: Edinburgh University Press.

Scott, M. (2011) *Wordsmith Tools*. Version 5.0. Oxford: Oxford University Press.

Stubbs, M. (1996). *Text and Corpus Analysis.* Oxford: Blackwell.

The methodological exploration of synergising CL and SFL in (critical) discourse studies: A case study on the discursive representation of *Chinese Dream*

Hang Su
University of Birmingham
suhangunique@hotmail.com

1 Introduction

This study presents a methodological exploration of combining corpus linguistics (CL; e.g. McEnery and Hardie 2012) and systemic functional linguistics (SFL; e.g. Halliday and Matthiessen 2004) in (critical) discourse studies. This exploratory process is demonstrated with a case study on the discursive representation of Chinese Dream which is put forth by the Chinese President Xi Jinping in 2012 and has been frequently discussed and reported in both Chinese and Western news media.

2 A brief literature review

There has been a long tradition of performing (critical) discourse analysis ((C)DA) within the framework of SFL, drawing particularly on concepts such as Transitivity, appraisal, and grammatical metaphor (e.g. Fowler 1991; Fairclough 1995, 2003; Young and Harrison 2004; Seo 2013; Hart 2014: 19-103). Recently, there also has been a growing interest of using corpus linguistic methods to facilitate discourse studies (e.g. Orpin 2005; Baker 2006, 2010; Baker *et al* 2008; Baker *et al* 2013a, b; Baker 2015). Overall, both SFL and CL have been shown to be influential and useful in (C)DA. However, this does not mean that the two approaches to (C)DA are not without critics. For example, while SFL provides a comprehensive explanatory framework for accounting for language use in social contexts, SFL is relatively less powerful to deal with a large amount of data; CL, on the other hand, enables the researcher to consider more data, but is relatively less capable of providing a more theoretical explanation of language use (cf. O'Donnell 2014). This indicates the necessity to explore how CL and SFL can be combined in (C)DA, which, however, has rarely been addressed.

In addition, the exploration of the discursive representation of a concept, a phenomenon or an event is one of the main areas that (C)DA is interested in. For example, Ricento (2003) looks into the discursive representation of *Americanism*, Powers and Xiao (2008) on SARS, Dunmire (2009) on 9/11 event, Alvaro (2013) on a dissident Liu Xiaobo, and Baker *et al.* (2013a) on Muslim. Apart from showing that language has an important role to play in social practice, these studies also suggest that it is feasible and worthwhile to explore how a concept, a phenomenon or an event is discursively represented from a linguistic perspective. Following this tradition, the present study, drawing on insights from CL and SFL, explores the discursive representation of the newly promoted concept *Chinese Dream*; the aims are 1) to show the complementarity between CL and SFL in (critical) discourse studies and 2) to offer a better understanding of the concept *Chinese Dream*.

3 Data and methodology

Nexis UK was used to collect the data. The source is restricted to a Chinese English-language news press, i.e. China Daily. The search term is specified as '*Chinese Dream* OR *China Dream*', and the time period from 1st January 2012 to 28th February 2015. The data is further restricted to those news texts where the search term '*Chinese Dream* OR *China Dream*' occurs at least three times in that period so as to make sure that the corpus compiled is highly representative of this topic. The retrieved texts are then cleared, excluding meta-information (e.g. author information, numbering of texts) and those similar texts, which gives me 142 texts. The corpus is thus compiled of the standardised 142 texts and contains 134,227 tokens.

Sketch Engine (Kilgarriff *et al* 2004) was used in the current study to perform the collocation analysis, to retrieve all the concordances containing 'Chinese Dream', and Word Sketch.

4 Analysis and discussion

This section reports the methodological exploration, including a collocation analysis, a systemic functional Transitivity analysis, and a corpus-assisted Transitivity analysis.

5 Collocation analysis

The analysis uses one of the typical corpus linguistic methods – the collocation analysis, to explore the discursive representation of Chinese Dream. Starting with the top 30 collocates of each major word class (i.e. noun, verb and adjective), the initial analysis suggests that the collocates of Chinese Dream can be generally categorised into four semantic or functional groups, i.e. Concept, Realisation, Aim, and Influence. The proportion each semantic group occupies is shown in Table 1. As table 1 shows, the collocation analysis shows that the discussion of *Chinese Dream* is primarily concerned with four aspects, i.e. the conceptualisation, the aim, the realisation and the influence. Though this gives us

an overview of the concept of *Chinese Dream*, there is something missing, that is, this does not tell us about how *Chinese Dream* is represented in terms of 'participation'.

Semantic category	Type	Token
Concept	66	1278
Realisation	10	169
Influence	8	95
Aim	6	91

Table 1: Semantic categories of the collocates

6 Transitivity analysis

As discussed above, the corpus analysis has not revealed how *Chinese Dream* is discursively construed in terms of participation. I thus draw on the systemic functional concept of Transitivity, attempting to explore what an SFL analysis can reveal about the discursive representation of *Chinese Dream*. Simply put, TRANSITIVITY in SFL is the grammatical system through which the world of experience is transformed into meaning, or more specifically, "into a manageable set of PROCESS TYPES" (Halliday and Matthiessen 2004: 170). Each process type makes "distinctive contributions to the construal of experience" (ibid: 174) and involves different participant roles; for example, Material process construes action and typically involves an Actor and a Goal, Relational process serves to characterise and typically involves Carrier/Identified and Attribute/Identifier.

The result of Transitivity analysis is shown in Table 2.

Process type	Participant roles	No.
Relational (288)	Ca./Idd[49].	243
	Att./Idr.	45
Material (213)	Goal	159
	Actor	54
Mental (63)	Phenomenon	61
	Senser	2
Verbal (37)	Verbiage	33
	Sayer	4

Table 2: Transitivity analysis

The result is largely consistent with the collocation analysis. For example, the result that the Relational process is the dominant process type that is used to construe *Chinese Dream* supports that the concept of *Chinese Dream* is frequently discussed, as relational process mainly serves "to characterise and to identify" (Halliday and Matthiessen 2004: 210); and the result that *Chinese Dream* is frequently construed as Goal in Material process supports that how *Chinese Dream* can be realised is also often discussed. It can thus be argued that Transitivity analysis is a useful tool for revealing the discursive representation of, for example, a concept in this study. However, though Transitivity analysis indeed gives us more details about how *Chinese Dream* is discursively represented, the manual analysis is time-consuming and work-intensive. This stimulates me to explore whether it is possible to use corpus methods to assist the Transitivity analysis.

7 A corpus-assisted Transitivity analysis

As manually annotating each instance containing *Chinese Dream* is laborious, I further explored the possibility of using corpus methods to assist the Transitivity analysis. The method used here is quite simple, that is, the Word Sketch in Sketch Engine. The basic function of word sketch is to provide "summaries of a word's grammatical and collocational behaviour" (Kilgarriff *et al* 2004). I mainly analysed the verbs provided by Word Sketch in terms of Transitivity, as transitivity is mainly concerned with verbs; the result is given in Table 3 below.

Process type	Participant roles	No.
Relational (211)	Ca./Idd.	198
	Att./Idr.	43
Material (158)	Goal	136
	Actor	22
Mental (30)	Phenomenon	30
	Senser	0
Verbal (21)	Verbiage	19
	Sayer	2

Table 3: A corpus-assisted Transitivity analysis

A glance at Table 2 and Table 3 would inform us that the results obtained through two different methods are highly reminiscent. For example, Relational process is predominant; *Chinese Dream* is typically construed as Carrier/Identified in Relational process and as Actor in Material process in both analysis. While the results are consistent, it has to be noted that the corpus-assisted Transitivity analysis is much easier to perform than the one discussed in Section 4.2. So, in general, it can be argued that the corpus-assisted Transitivity analysis can exploit the respective strengths, and at the same time, avoid the weaknesses, of collocation analysis and Transitivity analysis.

8 Conclusion

This study has mainly presented a methodological exploration of how CL and SFL can be combined in (C)DA. With a case study on the discursive representation of *Chinese Dream*, it has been shown that a corpus-assisted Transitivity analysis enables

[49] Ca. stands for Carrier, Idd. for Identified, Att. for Attribute, and Idr. for Identifier.

the researcher to deal with relatively easily a large amount of data, and allows the researcher to observe how a concept (e.g. *Chinese Dream*) is discursively constructed. Overall, it can be reasonably confident to conclude that it is possible (and worthwhile) to propose ultimately an integrated analytic framework for (C)DA which is corpus-assisted and SFL-informed.

Acknowledgements

This study is supported by China Scholarship Council (No. 2012[3024]) and The Ministry of Education, P. R. China (No. 14YJCZH148).

References

Alvaro J. 2013. Discursive representations of a dissident: The case of Liu Xiaobo in China's English press. *Discourse & Society* 24(3): 289-314.

Baker, P. 2006. *Using corpora in discourse analysis*. London: Continuum.

Baker, P. 2010. *Sociolinguistics and corpus linguistics*. Edinburg: Edinburg University Press.

Baker, P., et al. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society* 19(3): 273-306.

Baker, P., Gabrielatos, C. and McEnery, T. 2013a. Sketching Muslims: A corpus driven analysis of representations around the word 'Muslim' in the British press 1998-2009. *Applied Linguistics* 34(3): 255-278.

Baker, P., Gabrielatos, C. and McEnery, T. 2013b. *Discourse analysis and media attitudes*. Cambridge: CUP.

Baker, P. (ed.). 2015. Special issue in *Discourse & Communication* 9(2).

Dunmire, P. 2009. 9/11 changed everything: An intertextual analysis of the Bush doctrine. *Discourse & Society* 20(2): 195-222.

Fairclough, N. 1995. *Critical discourse analysis: The critical study of language*. London: Longman.

Fairclough, N. 2003. *Analysing discourse: Textual analysis for social research*. London: Routledge.

Halliday, M. A. K. and Matthiessen, C. M. I. 2004. *An introduction to functional grammar*. 3rd edition. London: Edward Arnold.

Hart, C. 2014. *Discourse, grammar and ideology: Functional and cognitive perspectives*. London: Bloomsbury.

Kilgarriff, A., et al. 2004. The sketch engine. In: Proceedings of Euralex, pp. 105-116.

McEnery, T and Hardie, A. 2012. *Corpus linguistics: Method, theory and practice*. Cambridge: CUP.

O'Donnell, M. 2014. "Systemic Functional Linguistics and Corpus Linguistics: Interconnections and current state". In Fang Yan & Jonathan Webster (eds.), *Developing systemic functional linguistics: Theory and application*, 345–369. London: Equinox.

Orpin, D. 2005. Corpus linguistics and critical discourse analysis: Examining the ideology of sleaze. *International Journal of Corpus Linguistics* 10(1): 37-61.

Ricento, T. 2003. The discursive construction of Americanism. *Discourse & Society* 14(5): 611-637.

Seo, S. 2013. Hallidayean transitivity analysis: The battle for Tripoli in the contrasting headline of two national newspapers. *Discourse & Society* 24(6): 774-791.

Twitter rape threats and the discourse of online misogyny (DOOM): From discourses to networks

Claire Hardaker
Lancaster University
c.hardaker@lancaster.ac.uk

Mark McGlashan
Lancaster University
m.mcglashan@lancaster.ac.uk

1 Introduction

This paper presents a selection of findings from an eighteen-month ESRC-funded project, entitled "Twitter rape threats and the discourse of online misogyny". This project was instigated by the events of summer 2013, when feminist campaigner and journalist, Caroline Criado-Perez was targeted with a sustained campaign of extreme, misogynistic abuse on Twitter. The abuse was triggered by Criado-Perez's petition challenging the Bank of England's decision to remove the image of Elizabeth Fry from the £5 note and replace it with that of Winston Churchill. The premise of the petition was to maintain the representation of influential women on British currency, since the appearance of men only could be deemed a "damaging message that no woman has done anything important enough to appear [on our banknotes]" (Criado-Perez, 2013).

The petition was successful and the Bank of England announced on the 24th of July 2013 that author Jane Austen's image will appear on the new £10 note issued in 2016. Following this announcement, Criado-Perez began receiving abuse through her Twitter account (@CCriadoPerez), including rape, death, and bomb threats. These threats broadened out to encompass several notable women, and were malicious and numerous enough to eventually warrant the prosecution of three individuals under §127 of the UK Communications Act (CPS, 2013).

This case, which featured prominently in the news for several weeks, placed investigative bodies, policy makers, and legislators under intense media scrutiny. Politicians, journalists, and targets alike called for improvements across the board, from site report-abuse functions to the prosecution of offenders. However, given the little empirical research into behaviours such as sending rape threats on Twitter, making evidenced, balanced, long-term management, policy, and legislation decisions is difficult.

2 Research questions

As a result of this, we secured an ESRC urgency grant to investigate this event from a linguistic perspective, with a range of research questions. Two of these are as follows:

- How are online threats made and what kinds of abuse do they involve?
    - What topics are salient in the making of rape threats?
    - What discourses are used as part of making online rape threats?
- Do abusers affiliate, i.e. do abusers belong to or create abusive online networks?
    - Do pre-existing networks of abusers exist?
    - Do abusers affiliate? If so, how?

3 Data

Within this paper, we identify a network of abusive users and consider the kinds of language and discourse communities those users adopt. The corpus in question comprises Twitter data that involves interactions of Caroline Criado-Perez's Twitter account, @CCriadoPerez. The sample is made up of three kinds of interactions – tweets (online posts made by users), mentions (tweets which include other account usernames), and retweets (tweets by one author which are reproduced by another user for their followers to see).

The sample spans ninety-two days of activity, from midnight 25/06/13 to midnight 25/09/13 inclusive. This period was selected by identifying the date that Criado-Perez first highlights an instance of abuse directed towards her (25/07/2013) regarding the successful Bank of England petition. This tweet effectively stands as "tweet zero" (from the medical parlance of "patient zero"—the first individual infected with a contagion that becomes an epidemic). Extrapolating outwards from this, a sample was taken for a full calendar month prior to this date to examine whether there was a history of abuse in the short term and for two full calendar months following this date to investigate how the abuse unfolded.

Aside from dates, additional sampling criteria were used to capture all instances of direct interaction occurring in relation to the @CCriadoPerez account, and this resulted in the Criado-Perez Complete Corpus (or CPCC). These criteria were all tweets by and to Criado-Perez, as well as all retweets by and of her. For the purposes of this study, however, less direct forms of interaction such as retweets were excluded from the CPCC. The results of this sampling procedure yielded the Criado-Perez Tweets & Mentions Corpus (henceforth, CPTMC) totalling 76,235 tweets.

For every kind of post made on Twitter, metadata is recorded which contains a number of attributes –

or properties – enabling a range of possibilities for analysis. These include screen name, username, the user's biography, the text of the tweet, and so forth. In this paper, we focus on analysing the Text attribute (though where relevant, data from other attributes has been retrieved throughout the analysis). To construct the CPTMC from the CPCC, the Text attribute was isolated, stripped of all hashtags, links, and mentions, and made readable for use with a concordance tool. This left a corpus of 76,235 tweets, totalling 1,014,222 words. For the purpose of answering the research questions, we used AntConc version 3.4.2m and Gephi 0.8.2 beta.

4 Scope

In the analysis, we implement methods from corpus linguistics to outline frequent topics of conversation occurring in the corpus. Whilst the findings from this analysis show that several topics and discursive/rhetorical strategies are highly frequent within the corpus, we focus primarily on talk relating to (sexually) aggressive behaviours. We begin our analysis by examining frequent features in the language of the CPTMC through examining a frequency wordlist. The frequent lexical items reveal a number of broadly identifiable topics (or discursive strategies) within the corpus, but due to limitations of space, we focus on the topics of (sexual) aggression and gender, as well as their intersections.

We also investigate whether communities form around these discourses, and whether (newly) distinguishable communities share in the production of certain discourses. We focus on constructions of rape and how different discourse communities form and construct themselves through shared linguistic practices and discourse vis-à-vis their discursive constructions of rape. We study two broad groups of Twitter users identified in the CPTMC corpus: high-risk and low-risk.

High-risk users were defined as Twitter profiles that contained evidence of: intent to cause fear of (sexual) harm; harassment; and potentially illegal behaviour. Low-risk users were defined as Twitter profiles that contained evidence of: offensive material; insults; ridicule; no (linguistic) evidence of intent to cause fear or threat of (sexual) harm; and spamming (as opposed to harassment). (For the sake of completeness, no-risk users were defined as Twitter profiles that contained evidence none of the above.)

A number of abusive users were pre-identified by Criado-Perez during the period covered within the data-sampling period. To track and identify more abusive users and their communicative networks, two methods of manual identification were employed. Users were identified through observing both directed connections (where a user mentions another in their tweet) and undirected or "ambient" connections whereby users might "simply be speaking about the same topic at the same time" (Zappavigna, 2014: 11). Both methods involved manual interpretation of the content of tweets and classification of users. Through repeating this process—following numerous directed and undirected connections—a total of 208 'risky' users were detected (147 low-risk, sixty-one high-risk).

Three separate subcorpora were created from the tweets of each user group, named CPTMC no-risk, CPTMC low-risk, and CPTMC high-risk. A keyword analysis was then conduced whereby both the CPTMC low-risk and CPTMC high-risk corpora were compared against the CPTMC no-risk corpus to assess differences in discourse between the user groups and to assess whether different discourse communities exist.

Several frequent keywords were shared by low- and high-risk users in the CPTMC suggesting an interface between language and discourse with regards to sexual violence (*rape, raep*) and misogynistic insults (*bitch, cunt*) that may be characteristic of risky users engaged in making or talking about rape threats. However, whilst this mutual interest in similar lexis may indicate that they are part of a wider discourse community, differences between the groups also exist.

Finally, we present some visual networks of these low- and high-risk users to represent how such networks form, and how they function.

In short, a larger, nebulous discourse community emerged from the analysis, and within this, it was possible to identify a smaller community of low-risk users (those who tweeted insults and sarcasm), and a smaller-still community of low- and high-risk users (those who tweeted threats, harassment, and even breached any number of UK laws).

It would be easy to automatically discount the low-risk users from their place in the larger community, however, it is worth considering that similarities between the discourses shared by these communities could facilitate a user's gradual escalation from low-risk (unpleasant) through to high-risk (illegal) online interaction, possibly without even being quite aware of that gradual shift. Indeed, both the low- and high-risk abusers coalesced not only around the discussion of rape, but also of homophobia and racism.

Acknowledgements

This work was supported by the Economic and Social Research Council [grant number ES/L008874/1].

References

CPS. 2013. 'CPS authorises charges in Twitter-related cases.' no. Available online at http://www.cps.gov.uk/news/latest_news/stella_creasy_mp_caroline_criado-perez/ (Accessed 30 September 2014)

Criado-Perez, C. 2013. 'We Need Women on British Banknotes.' Available online at http://www.change.org/en-GB/petitions/we-need-women-on-british-banknotes (Accessed 10 August 2013)

Zappavigna, M. 2014. 'Enacting identity in microblogging through ambient affiliation.' *Discourse & Communication* no. 8 (2): 209-228.

Employing Learner Corpus in EAP Classes: The METU TEEC Example

Çiler Hatipoğlu
Middle East Technical University
ciler@metu.edu.tr

Yasemin Bayyurt
Boğaziçi Univesity
bayyurty@boun.edu.tr

English is an essential component of all levels of national education in Turkey and now it is "the most commonly taught foreign language in Turkish schools" (Bayyurt 2010:163). There are both state and private schools and universities in Turkey where the medium of instruction is English as well as hundreds of private language courses, and their number is getting bigger every year (Hatipoğlu 2013). Despite the popularity of English in the country and "despite the time, money and effort spent on foreign language education in Turkey, low foreign language proficiency level has remained a serious problem" (Işık 2008:15). Practitioners and researchers alike agree that "English language teaching/learning is problematic in Turkey" (Kızıldağ 2009:189). This is why, the last two decades have seen spike in research aiming to uncover the reasons behind these problems (Aktas 2005; Oguz 1999; Paker 2007; Şallı-Çopur 2008; Tilfarlioglu & Ozturk 2007). Among the various determinants (e.g., language planning, foreign language teaching methodologies, student interest and motivation) one comes to the forefront in those discussions, that is, the language teacher and his/her knowledge of the target language. The language teacher is an important variable determining the success or failure of the language teaching process in Turkey since English is a foreign language in the country and it has no official status beyond the 'classroom walls' of English medium institutions (Bayyurt 2012; Doğançay-Aktuna 1998). Students in Turkey "typically receive input in the new language only in the classroom" (Oxford & Shearin 1994:14) and during the majority of the day they are surrounded by their mother tongue. This, in turn, means that the teachers are role models for their students and their knowledge and skills in the foreign language frequently determine whether their students would become motivated and successful language learners and skilful communicators in the target language or not (Hatipoğlu 2013).

Taking into consideration the importance of the language competence of language teachers for successful language teaching in the country and the problems related to teaching English in Turkey, it was decided to embark on a project aiming to, first, create a specialised corpus of academic English

including samples coming from non-native pre-service English language teachers in Turkey and, then to use this learner corpus both in creating teaching materials and in implementing data driven learning in undergraduate courses which are part of the curriculum of the English Language Teacher training programs in Turkey.

The corpus created for this project is a specialised Turkish English Exam Corpus (TEEC), which has been compiled by a research team at Middle East Technical University (METU), Ankara, Turkey. The corpus consists of 1914 Linguistic and ELT exam papers (955483 words) written in timed circumstances with no access to reference materials by the students at the Foreign Language Education (FLE) department at METU, Ankara between January 2005 and December 2012. Only exam papers were included in the corpus since the aim was to collect spontaneous data which are the more realistic representations of the English of the pre-service English language teachers (Ellis 2001; Selinker 1972). Since the aim in creating this corpus was to identify the characteristics of the English used by pre-service English language teachers (who were also advanced learners of English) it was decided that the corpus would be more useful to potential users if it were tagged for features such as orthography, punctuation, grammar (i.e., word formation, agreement, tense, mood, word order) as well as discoursal, pragmatic and rhetorical characteristics. The annotation in the corpus was done using EXMARaLDA partiture editor (i.e., Extensive Markup Language for Discourse Annotation; http://exmaralda. org/).

The analyses of the mistakes of the learners comprised three stages: (1) identification and isolation, (2) supplying the target form and (3) classification of the problem. The classification of the identified problems, on the other hand, was done using the scheme devised by Dulay et al. (1982) and it included categories such as "*omission", "addition", "misinformation"* and *"misordering"*.

The first course where the usefulness of the METU TEEC was tested was *The English Lexicon* (TEL) course. TEL is one of the must courses in the curriculum of the English Language Teaching Programs in Turkey and its main goal is to present, discuss and analyze topics that are difficult for native speakers of Turkish learning English. By focusing on those problematic topics the course aims not only to equip students with tools that will help them do in-depth analyses of the linguistic data coming from non-native speakers of English but also to assist them in improving their own knowledge of the target language. Since the compilation of the METU TEEC, the topics included in the course outline and the teaching methodology followed in the classes have been based on the analysis of the METU TEEC.

So, this paper will first, present specific examples of how the data in METU TEEC were tagged and analysed and, then, will, discuss how a fourth year course entitled *The English Lexicon* was structured as a result of those analyses. Finally, the pluses and minuses of creating a corpus-informed course from the point of view of both learners and instructors will be discussed.

References

Aktas, T. 2005. Yabanci Dil Ogretiminde Iletisimsel Yeti. *Journal of Language and Linguistic Studies*, 1(1), 89-100.

Bayyurt, Y. 2010. Author positioning in academic writing. In S. Zyngier and V. Viana (Eds), *Avaliaçoes E Perspectivas: Mapeando Os Estudos Empiricos Na Area deHumanas (Appraisals and Perspectives: Mapping Empirical Studies In The Humanities)* (pp. 163-184). Rio de Janeiro: The Federal University of Rio de Janeiro.

Bayyurt, Y**.** 2012. Proposing a model for English language education in the Turkish socio-cultural context. In Yasemin Bayyurt and Yeşim Bektaş-Çetinkaya (Eds.), *Research Perspectives on Teaching and Learning English in Turkey: Policies and Practices* (pp. 301-312). Berlin: Peter Lang.

Doğançay-Aktuna, S. 1998. The spread of English in Turkey and its current sociolinguistic profile. *Journal of Multilingual and Multicultural Development*, 19 (1), 23-39.

Dulay, H. C., Burt, M. K. and Krashen, S. 1982. *Language Two.* New York: Oxford University Press.

Ellis, R. 2001. *The Study of Second Language Acquisition.* Oxford: Oxford University Press.

Hatipoğlu, Ç. 2010. Summative Evolution of an Undergraduate 'English Language Testing and Evaluation' Course by Future English Language Teachers. *English Language Teacher Education and Development (ELTED),* 13 (Winter 2010), 40-51.

Hatipoğlu, Ç. 2013. First Stage in The Construction Of METU Turkish English Exam Corpus (METU TEEC). *Boğaziçi University Journal of Education*, 30 (1), 5-23.

Isik, A. 2008. Yabanci Dil Egitimimizdeki Yanlislar Nereden Kaynaklaniyor? *Journal of Language and Linguistics*, 4(2), 15-26.

Kızıldağ, A. 2009. Teaching English in Turkey: Dialogues with teachers about the challenges in public primary schools. *International Electronic Journal of Elementary Education*, 1 (3), 188-201.

Oguz, E. 1999. İlköğretimde Yabancı Dil (İngilizce) Öğretimi Sorunları (The Problems of foreign language (English) teaching in elementary schools). Unpublished Master Thesis. Kocaeli University: Kocaeli, Turkey.

Oxford, R. and Shearin, J. 1994. Language Learning Motivation: Expanding the Theoretical Framework. *The Modern Language Journal*, 78 (1), 12-28.

Paker, T. 2007. Problems of teaching English in schools in Çal Region and suggested solutions. 21. Yüzyıla Girerken Geçmişten Günümüze Çal Yöresi: Baklan, Çal, Bekilli. Çal Yöresi Yardımlaşma ve Dayanışma Derneği Yayını, 3, 684-690.

Şallı-Çopur, D. 2008. Teacher Effectiveness In Initial Years Of Service: A Case Study On The Graduates Of METU Foreign Language Education Program. Unpublıshed PhD Thesis. Middle East Technical University: Ankara, Turkey.

Selinker, L. 1972. Interlanguage. *IRAL,* 10 (3), 209-231.

Construction of male and female identities by a misogynistic murderer: a corpus-based discourse analysis of Elliot Rodger's manifesto

Abi Hawtin
LancasterUniversity
a.hawtin@lancaster.ac.uk

1 Background

On 23rd May 2014, 22 year old Elliot Rodger killed 6 people and injured 13 others in California. He left behind a series of YouTube videos in which he discussed his hatred of women, and a 'manifesto' which outlined his life up to that point, his views on women, and his plan to take revenge[50]. In this study I use corpus methods (collocation and semantic collocation analysis) to analyse the ways in which Rodger constructs the identities of men and women in his manifesto, to investigate whether the way he views men and women represents a new and more dangerous type of misogyny than has previously been studied in detail.

Corpus methods have long been used to analyse representations of gender in language; for instance, Pearce (2008) finds that in the BNC men are often in an active position and women a passive position, and Herdağdelen and Baroni (2011) find the same in ukWaC (Ferraresi et al, 2008), with men most often discussed relative to positions of power and women in terms of having children. Furthermore, Caldas-Coulthard and Moon (2010) find that men are evaluated in terms of social status and behaviours, whereas women are most often evaluated in terms of appearance. Despite a wealth of research into gender representation in general discourse, there has been little corpus-based research into explicitly misogynistic texts, even though much recent research suggests that misogynistic views are becoming normalised (Jane, 2014; Horvarth et al, 2012). This study of Rodger's manifesto (an extreme and violently misogynistic text) addresses that gap in the research to date.

Rodger is perhaps unusual in that he wrote at sufficient length prior to his murders that the resulting document ('My Twisted World: The Story of Elliot Rodger'[51]) is large enough for corpus analysis by itself. Although commonly referred to as a 'manifesto', it contains an autobiographical account with a particular focus on his relationship with women and his plans to punish them.

[50] http://www.bbc.co.uk/news/world-us-canada-27562917
[51] http://abclocal.go.com/three/kabc/kabc/My-Twisted-World.pdf

2 Methodology

I approach this investigation by looking at statistical collocations – both at the level of the word, and at the level of the semantic tag. This method has been productively employed by Rayson (2008) in a similar analysis. To look at collocations, it is necessary to first define the node(s) one is examining. My aim is to investigate how men and women are represented in the manifesto, but simply searching for the word-forms 'men' and 'women' would not uncover all of Rodger's references to men and women. To select appropriate search terms, I loaded the manifesto into Wmatrix (Rayson, 2009), which incorporates the USAS sematic tagger (Wilson and Rayson, 1993) and then searched for all words tagged as 'S2.1' (People: female) or 'S2.2' (People: male). This gave me a list of terms tagged as referring to female and male persons. Not all these terms were relevant to my analysis; most notably, a number refer to specific individuals (e.g. Rodger's family members) and are thus not relevant to understanding how Rodger constructs the identities of men and women *in general*. I thus compiled a list of search terms based on the 'S2.1/S2.2' result, but not including any such individual-reference terms. The final list of search terms whose collocations I went on to analyse is as follows:

S2.1 – People: female	S2.2 – People: male
Female	Male
Females	Males
Woman	Man
Women	Men
Girl	Boy
Girls	Boys
Girlfriend	Guy
Girlfriends	Guys
	Boyfriend
	Boyfriends

Table 1: Search terms for collocation analysis.

The first type of analysis I performed was a collocation analysis, using AntConc (Anthony, 2014). I searched for collocates of all of the search terms listed above, with a minimum frequency of 5 and a span of 5 words left and right. Collocations were only considered in the analysis if they had an MI of 3 or higher (Hunston, 2002: 71-72; Durrant and Doherty, 2010: 145). I subsequently used Wmatrix (Rayson, 2009) to conduct a similar collocation analysis, but using the USAS tags and exploiting Wmatrix's ability to search for semantic tag collocates.

3 Analysis

Looking at the collocates of the female and male search terms helps to reveal several discourses which contribute to the ways that Rodger constructs the identities of men and women in his manifesto. Primarily, he constructs *both* men and women (i) as homogenous groups which he is outside of, and (ii) in terms of their appearance. Rodger furthermore constructs women as objects of his hatred, targets for his revenge, as goals which he wants to achieve, but also as having power over him and men in general. By contrast, he constructs men as a group as being able to have experiences, usually sexual, which he cannot. Table 2 gives some examples of the collocates which contribute to these discourses.

Collocate	Collocates with:	MI value
Humanity	women	8.53
Pretty	girl(s)	8.43
Hot	girl(s)	8.36
beautiful	girlfriend(s)	8.30
Blonde	girl(s)	7.74
Against	women	7.71
Hatred	women	7.21
Worthy	girl(s)	6.90
sexual	girl(s)	6.56
Love	girl(s)	6.45
Virginity	girl(s)	6.29
Experience	girlfriend(s)	6.27
Sex	girl(s)	6.23
all	women	5.40
Black	boy(s)	8.60
Other	men	7.65
Able	boy(s)	6.83
Experience	men	6.49

Table 2: Collocates of the male and female search terms.

The semantic collocates found for women were 'Disease', 'Undeserving', 'Unwanted', and 'Relationship: Intimacy and sex'. 'Disease' words, primarily *pain*, are used both to express Rodger's feeling that women have caused him pain and also to refer to his plans to inflict pain upon *them*. This 'pain' discourse is distinct from, but in Rodgers' understanding a direct result of, the state-of-affairs represented via the remaining 3 semantic tag collocates. All these tags contribute to the expression of a discourse of women having the power to 'choose' men. However, it becomes clear when looking at the concordance lines (see concordance lines below) that Rodger feels that women are using this power wrongly by choosing the 'wrong' men.

- to this filthy scum, but they **reject** ME? The injustice! Females truly

- those evil, slutty bitches who **rejected** me, along with the fraternity
- never have love. Girls deem me **unworthy** of it, I thought myself over
- that no girl in the world wanted to **fuck** me. I was a kissless virgin after
- teenagers, I never had my first **kiss**, I never held hands with a girl
- way home. Why does he deserve the **love** of a beautiful girl, and not me?

The semantic tag collocates of the male search terms also contributed to this discourse of women as unjustly powerful (see concordance lines below). The semantic tags 'Colour and colour patterns' and 'Judgment of appearance: beautiful' reveal his feelings that only men with certain appearances deserve to be chosen by girls. The 'Colour and colour patterns' tag reveals an instance of mistagging where 'black' has been tagged as a description of colour rather than race. This, however, leads to evidence of a discourse of racism, where 'ugly' frequently occurs with 'black' (see concordance lines below). The semantic tag 'Able/Intelligent' is used to express frustration that even men with the 'wrong' appearance or personality are 'able' to be chosen by girls and to have experiences with them that Rodger cannot.

- How could an inferior, ugly **black** boy be able to get a white girl and
- is actually true, if this ugly **black** filth was able to have sex with a
- do it alone while other men were **able** to do it with their girlfriends
- The short, chubby guy was **able** to get a girl into his room before I

This *unjust power* discourse is by far the most prominent element of Rodger's representations of men and women, with almost all of the collocations (with both words and semantic tags) contributing in some way towards his construction of women as having a great amount of power, but using this power wrongly.

4 Conclusion

The analysis above shows that Elliot Rodger consistently expresses and constructs an ideology that in contemporary media commentary is commonly called the 'new misogyny' (see, for instance, Marcotte 2014). This is a worldview in which men view women as privileged and powerful, and themselves as oppressed. This kind of sexism stands in contrast to that discussed in the earlier research which I reviewed above – misogyny which usually represents men in positions of power and women as less powerful, i.e. a more traditional patriarchal ideology. This is in contrast to Rodger's construction of the relationship between men and women, in which he clearly shows that he believes that women hold all power over men. That, I suggest, is the key difference between Rodger's (ultimately murderous) views and 'everyday' sexist views rooted in traditional patriarchy. Further study is needed to establish how much of this worldview is specific to Rodger and how much is generally characteristic of the 'new misogyny'.

References

Anthony, L. 2014. AntConc (Version 3.4.3w) [Computer Software]. Tokyo, Japan: Waseda University. Available at: http://www.laurenceanthony.net/

Caldas-Coulthard, C. & Moon, R. 2010. "Curvy, Hunky, Kinky": Using corpora as tools for critical analysis. *Discourse and Society,* 21(2), pp. 99-133.

Durrant, P. & Doherty, A. 2010. Are high frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory, 6*(2), pp. 125- 155.

Ferraresi, A., Zanchetta, E., Baroni, M., & Bernardini, S. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the WAC4 Workshop at LREC 2008*, Marrakech, Morocco.

Herdağdelen, A. & Baroni, M. 2011. Stereotypical gender actions can be extracted from web text. *Journal of the American Society for Information Science and Technology, 62*(9), pp.1741-1749.

Horvarth, M., Hegarty, P., Tyler, S. & Mansfield, S. 2012. "Lights on at the end of the party": Are lads' mags mainstreaming dangerous sexism? *British Journal of Psychology, 103,* pp. 454-471.

Hunston, S. 2002. *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

Jane, E. 2014. 'Back to the kitchen, cunt': speaking the unspeakable about online misogyny. *Continuum, 28*(4), pp. 558-570.

Marcotte, A. 2014, May 30. *4 myths about sex and women that prop up the new misogyny.* Retrieved from:http://www.salon.com/2014/05/30/4_myths_about_sex_and_women_that_prop_up_the_new_misogyny_partner/

Pearce, M. 2008. Investigating the collocational behaviour of MAN and WOMAN in the BNC using Sketch Engine. *Corpora, 3*(1), pp. 1-29.

Rayson, P. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics, 13*(4), pp. 519-549.

Rayson, P**.** 2009. Wmatrix: a web-based corpus processing environment, Computing Department, Lancaster University. Available at: http://ucrel.lancs.ac.uk/wmatrix/

Wilson, A. & Rayson, P. 1993. Automatic content

analysis of spoken discourse. In: C. Souter & E. Atwell (Eds.). *Corpus based computational linguistics,* (pp. 215-226), Amsterdam: Rodopi.

Investigating collocation using EEG

Jennifer Hughes
Lancaster University
j.j.hughes@lancaster.ac.uk

1 Introduction

Previous studies have investigated the processing of collocation using self-paced reading and eye-tracking experiments (e.g. McDonald and Shillcock 2003a, 2003b; Millar 2010). However, to date, there are no studies which investigate the processing of collocation using electroencephalography (EEG). I will address this methodological research gap by discussing how EEG can be used to determine whether or not collocation is a psychologically observable phenomenon, in terms of whether or not collocation strength as measured in text corpora can accurately predict the manner and speed at which language is processed.

2 Prior work on the psycholinguistics of collocation

Millar (2010) conducts two self-paced reading experiments and one eye-tracking experiment and finds that learner collocations are associated with an increased processing burden in comparison to their native speaker equivalents. Learner collocations are operationalized by Millar (2010) as collocations that occur in a learner corpus but do not occur in the BNC and are intuitively unacceptable from a native speaker's perspective; the native speaker equivalents are collocations with an equivalent meaning and a mutual information (MI: the binary logarithm of the ratio of observed to expected frequency of the collocation) score above 3 in the BNC.

In an additional self-paced reading experiment, Millar (2010) finds an increased processing burden for collocations that do not represent errors but are weak (with an MI of below 3) in comparison to semantically equivalent collocations (with an MI of above 8). The results from these studies suggest that stronger collocations have a processing advantage over weaker collocations and collocation errors.

Further evidence demonstrating that the psychological underpinnings of collocation can be experimentally validated comes from McDonald and Shillock's (2003a, 2003b) eye-tracking experiments. Focusing on the transitional probabilities (i.e. the probability of word B being produced given that the previous word was A), McDonald and Shillcock (2003a, 2003b) find that participants fixate on the second word of a bigram for a significantly longer period of time if they are reading a bigram with a low transitional probability (e.g. *avoid discovery*)

compared to a bigram with a high transitional probability (e.g. *avoid confusion*).

Furthermore, in a more recent study, Huang et al. (2010) conduct an eye-tracking experiment and find that both native speakers and learners of English process the final word of a sequence significantly faster if the word is part of a sequence that has high transitional probabilities than if the word is part of a sequence that has lower transitional probabilities. The results of these eye-tracking experiments suggest that collocations with a high transitional probability can be processed more quickly, supporting the notion that psychological links exist between collocation pairs.

3 What is EEG and how can it be used in the study of collocation?

EEG is "a means of measuring electrical potentials in the brain by placing electrodes across the scalp" (Harley 2008). EEG recording devices measure event-related potentials (ERPs) (Ashcraft and Radvansky 2010), i.e. "the momentary changes in electrical activity of the brain when a particular stimulus is presented to a person" (Ashcraft and Radvansky 2010).

EEG has already been used in the study of language. For example, previous studies have shown that sentences which contain semantic errors (Kutas and Hillyard 1980) and syntactic errors (Osterhout and Holcomb 1992) elicit distinct ERP responses. However, no studies have investigated the ERP responses elicited by sentences which contain collocation errors in the absence of other semantic or syntactic errors.

When using EEG to investigate the processing of collocation, two stimuli conditions are presented to native speakers. Condition 1 contains stronger collocations with a higher MI and transitional probability; condition 2 contains either collocation errors or weaker collocations with a lower MI and transitional probability. If the ERP responses are found to be different when native speakers are processing the stronger collocations compared to the weaker collocations or collocation errors, this will provide strong support for the idea that detectable neural/psychological processes underlie the phenomenon of collocation in language usage.

The items in both conditions are controlled for length and frequency, and are embedded into sentences that are equally plausible in the sense that they depict ''an entity or situation that is likely to occur in the real world" (Arnon & Snider 2010). The preceding contexts are identical in both conditions and create an equally "low contextual constraint" (Millar 2010:108), meaning that the collocations are not primed by the semantics of the preceding words.

Some participants are asked to read the sentences in order to investigate whether or not the ERP response is different when native speakers read stronger collocations as compared to weaker collocations/collocation errors; other participants are asked to listen to a recording of another native speaker reading the sentences aloud in order to investigate whether or not the ERP response is different when native speakers listen to stronger collocations as compared to weaker collocations/collocation errors.

After one sentence is presented, the next sentence is automatically presented to participants. This is because, if the participants had to press a button in order for the next sentence to be revealed, the muscle activity involved in this action would stimulate activity in the motor cortex which, in turn, would affect the EEG data.

4 How is the EEG data interpreted?

When interpreting the EEG data, the aim is to look for evidence of processing differences between stronger collocations and weaker collocations or collocation errors. These processing differences can relate to the manner or the speed at which the language is processed. For instance, it could be the case that the weaker collocations are processed more slowly than the stronger collocations, thereby increasing the duration of the ERP response. It could also be the case that a certain ERP component is engaged to a different degree across the two conditions, or the ERP activity could be distributed differently across the scalp (Otten and Rugg 2005). Any of these differences would demonstrate that collocation strength as measured in text corpora can accurately predict the manner or speed at which language is processed, thereby providing strong support for the idea that detectable neural/psychological processes underlie the phenomenon of collocation in language usage.

References

Arnon, I. and Snider, N. 2010. "More than words: Frequency effects for multi-word phrases". *Journal of Memory and Language* 62 (1): 67-82.

Ashcraft, M. H. and Radvansky, G. A. 2010. *Cognition* (5th edn.). London: Pearson.

Harley, T. A. 2008. *The psychology of language: From data to theory* (3rd edn.). New York: Psychology Press.

Huang, P., Y. Wible, D., and Ko, H. W. 2012. Frequency effects and transitional probabilities in L1 and L2 speakers' processing of multiword expressions. In S. Th. Gries and D. Divjak (eds.). *Frequency effects in language learning and processing*. Berlin: De Gruyter Mouton.

Kutas, M. and Hillyard, S. A. 1980. “Reading senseless sentences: Brain potentials reflect semantic incongruity”. *Science* 207 (4427): 203-205.

McDonald, S. A. and Shillcock, R. C. 2003a. “Eye movements reveal the on-line computation of lexical probabilities during reading”. *Psychological Science* 14 (6): 648-652.

McDonald, S. A. and Shillcock, R. C. 2003b. “Low-level predictive inference in reading: the influence of transitional probabilities on eye movements”. *Vision Research* 43 (16): 1735-51.

Millar, N. (2010). *The processing of learner collocations* Unpublished PhD thesis. Lancaster University.

Osterhout, L. and Holcomb, P. J. 1992. “Event-related brain potentials elicited by syntactic anomaly”. *Journal of Memory and Language* 31 (6): 785-806.

Otten, L. J. and Rugg, M. D. 2005. “Interpreting event-related brain potentials”. In T. C. Handy (ed.). *Event-related potentials: A methods handbook*. Cambridge, MA: MIT Press.

CSAE@R: Constructing an online monitor corpus of South African English

Sally Hunt
Rhodes University
s.hunt@ru.ac.za

Richard Bowker
Rhodes University
r.bowker@ru.ac.za

In 2014 we completed the first stage of the automated build of a large monitor corpus of South African English (SAE). This corpus, of online South African media texts in English, was compiled using custom-built semi-automated collection software. We have focussed in our analysis of the data on two aspects of SAE: trends in terms of the linguistic origins of borrowed terms, and the use of selected modals in SAE.

The construction of the software system takes its lead from the design of a similar kind of corpus, the Norwegian Newspaper Corpus, which was designed to identify neologisms in Norwegian, as reported by Andersen (2011). Kilgarriff et al.’s (2006) development of the New Corpus for Ireland has also been influential. The first component of the CSAE@R build is the piping together of software modules that crawl specific sites on the web (up to now, limited to South African media sites), access relevant material, extract metadata, remove html links and de-duplicate the data, and save the texts, along with the metadata, to a database. This is intended to be analysable via an external concordancer, which is the focus of our project for 2015. The initial collection has resulted in a corpus of approximately 70 million words of contemporary media texts published online in English in South Africa, which itself is a substantial amount of data to answer questions about SAE. Once the software begins its monitor function, sourcing new texts as they are published, and working back into the archives, it should prove to be particularly useful for lexicographic purposes, processing approximately 2.4 million words per month, with data taken from a broader range of sources such as online newspapers, magazines, blogs, and including the online comments sections of these texts where available. In terms of tracing the movement of lexical items into the language variety, and shifts in usage on the level of discourse and ideology, CSAE@R will be indispensable and the first large corpus of its kind in the region.

Although there have been corpora constructed, or partially constructed, with South African English data, such as De Klerk's (2006) corpus of spoken Xhosa English, and Jeffrey's ICE-SA corpus which currently stands at 554 810 words (Wasserman and Van Rooy 2014), to date, no comparable corpus for

South African English (SAE) exists. Van Rooy and his colleagues (see Rossouw and Van Rooy (2012) and Wasserman and Van Rooy (2014) for example) have constructed a corpus of historical White South African English from 1820 to 1990 totalling 123 247 words, which they have used in conjunction with ICE-SA (for more contemporary texts), to explore various features of SAE. However, even in combination, the data comprise a fairly small corpus of a narrow, ethnically defined, variety, although the collection does benefit from a variety of text types.

SAE11 (Hunt and Bowker 2013), a one-million word snapshot corpus of SAE using the Brown-LOB sampling frame, informed the current project both methodologically and in terms of suggesting avenues for linguistic research that might be productive with a larger corpus. Chief among these are lexicographic applications which undoubtedly benefit from a larger data set (see Krishnamurthy 2000). This is especially the case for SAE, which comprises a relatively small percentage of the lexical items used in everyday English in South Africa. A second main application is the investigation of various features of SAE, including syntactic and pragmatic features, and aspects of discourse, which have not had the benefit of a large corpus for confirmation or elimination. The focus currently in our work is on the patterns revealed in terms of modality.

The design of the software system, and how it was influenced by the Andersen, Kilgariff et al. and other models, as well as the problems we have needed to overcome in order to implement the system, are worthy of a more detailed discussion. We would also like to suggest means by which the system can be made to function to identify previously unrecorded lexical items in SAE.

The project, by developing a monitor corpus, is a way of querying Hanks' (2010) contention that corpora are not an especially useful means of identifying new lexical items for lexicographic purposes. The most productive source of neologisms in SAE is borrowing from other local languages. Conversely, the presence of other languages, both Bantu and Indo-European, in South Africa, inevitably affects the English variety, and that influence is felt predominantly lexically. Moreover, there are discernible periods in the history of the country in which some languages are more prominent than others in terms of lexical donation. Early loan words are mostly from the indigenous Khoi and San languages e.g. *dagga* (cannabis, first recorded in SAE in 1670) and Cape Dutch/Afrikaans e.g. *commando* (a military unit, 1790), due to the social contact between these groups and British English speakers who arrived at the Cape. There is a long and rich tradition of borrowings from Afrikaans, in particular, and the use of these terms is not restricted to occasional substitution: they are (currently) the only SAE terms for their referents, to the extent that Lass (1995: 382) suggests that "many English speakers do not know other words for some of these items: *bakkie* (light delivery van), *bergie* (vagrant), *braai* (barbeque), *dassie* (hyrax), *erf* (plot of land), *kloof* (ravine), *nogal* (what's more), *ou/oke* (bloke, chap)". The corpus is able to show the assimilation of borrowed items in terms of the grammatical affixes in cases where these vary between English and Afrikaans. However there are many instances in which the plural markers, for instance, for English and Afrikaans nouns would be identical, so the assimilation is visible only in terms of pronunciation, something which we cannot access in the written corpus.

Later borrowings, however, appear to be increasingly from the Nguni group of languages: mainly from isiZulu and isiXhosa e.g. *izinyoka* (literally *snakes*, meaning *electricity thieves*); and from the Sotho group of languages: predominantly Sesotho and Setswana e.g. *kwerekwere* (a pejorative term for a non-South African African person). This may be as a result of the political power and social prestige associated with the speakers of these languages, or the nature of the contact between speakers of African languages and speakers of English. Indeed, the rise in the number of speakers of English from racial groups other than white, as well as the increase in second language speakers of English (SA Census 2011), might account for the increase in borrowed terms. Of course, many of the original borrowings continue to be used, but new loan words are more likely to be Nguni or Sotho in origin, rather than Khoi/San or Afrikaans, due to the near extinction of the former and the political eclipse of the latter. Borrowings from African languages display varying degrees of assimilation in terms of spelling and phonological adaptation which makes their identification in the data especially complicated, especially as in these languages grammatical affixes are found at the beginning of words, while lexical items which have undergone significant assimilation may take English suffixes instead. In addition, cognate lexical items may be donated by several related languages, resulting in similar, but not identical, orthographies. Lemmatising SAE words is a particular challenge for the analysis of this data.

In terms of our second focus, that of evidence for or against claimed features of SAE, we turn to the use of modals in SAE. Of particular interest is the use of *must* as non-obligative in SAE, more equivalent to *shall* or *should*, probably from Afrikaans 'moet', which is a far weaker verb (cf. Jeffrey and Van Rooy 2004, Wasserman and Van

Rooy 2014). Unlike the corpus built by Van Rooy and his colleagues, who used historical texts produced exclusively by White speakers of SAE, CSAE@R includes the English used by South Africans from a greater variety of linguistic and ethnic backgrounds, which enables us to identify trends across the variety as a whole, as well as other, more fine-grained patterns not as clearly evident in a smaller corpus. Both SAE11 and CSAE@R provide firm evidence for this usage, across a range of genres, including more formal text types.

References

Andersen, G. 2011. "Corpora as lexicographical basis – The case of anglicisms in Norwegian". *Studies in Variation, Contacts and Change in English* (VARIENG) 6. http://www.helsinki.fi/varieng/journal/volumes/06/andersen/

Kilgarriff, A., Rundell, M. & Uí Dhonnchadha, E. 2006. "Efficient Corpus Development for Lexicography: Building the New Corpus for Ireland" in *Language Resources and Evaluation* 40(2): 127-152 .

Hanks, P. 2010. "Compiling a monolingual dictionary for native speakers" in *Lexikos* 20: 580-598.

Hunt, S.A. and Bowker, R. 2013. "SAE11: a new member of the family". Paper presented at Corpus Linguistics 2013 at Lancaster, UK, 22 - 26 July 2013.

Jeffrey, C. and Van Rooy, B. 2004. "Emphasiser now in colloquial South African English". *World Englishes* 23(2): 269-280.

Krishnamurthy, R. 2000. "Size matters: Creating dictionaries from the world's largest corpus" in *Proceedings of KOTESOL 2000: Casting the Net: Diversity in Language Learning*, Taegu, Korea: 169-180.

Lass, R. 1995. "South African English". In R. Mesthrie, ed. *Language and Social History: Studies in South African Sociolinguistics*. Cape Town: David Phillip, 89-106.

Rossouw, R. & Van Rooy, B. 2012. "Diachronic changes in modality in South African English". *English World-Wide* 33(1): 1-26.

Wasserman, R. and Van Rooy, B. 2014. "The Development of Modals of Obligation and Necessity in White South African English through Contact with Afrikaans". *Journal of English Linguistics* 42(1): 31–50.

A text analysis by the use of frequent multi-word sequences: D. H. Lawrence's *Lady Chatterley's Lover*

Reiko Ikeo
Senshu University
rikeo0919@gmail.com

Multi-word sequences which occur frequently in a text can often be seen to play a particular textual function of evaluation and present information management. In fiction, some fixed phrases associated with particular protagonists in texts may be noticeable to the reader, and this helps to characterise the protagonists. Other multi-word phrases, on the other hand, are frequent but less salient, and thus, only the n-gram function of a corpus concordancer program can identify them. Their textual functions need to be examined in the context in which the expressions occur, comparing one example with another within the text, or in multiple texts by the same author as manifesting the authorial stylistic traits.

This paper examines how particular multi-word sequences and a set of adjectives which are closely related to the leading protagonists' viewpoints contribute to the character development and narrative construction in the fictional text, D. H. Lawrence's *Lady Chatterley's Lover* (LCL) (1960 [1928]). LCL is an iconic novel which explores sensuality and sexuality as an essential part of humanity. Nouns of sexual organs and gender-related nouns of body parts are found in two keyword lists, which were respectively made by comparing the text with the fiction section of the BLOB 1931 Corpus and with my own collection of Lawrence's six novels taken from Project Gutenberg *(England, My England*, *The Rainbow*, *Sons and Lovers*, *Tresspasser*, *The White Peacock* and *Women in Love*). The romance between the mistress of the house of an aristocratic family and the family's gamekeeper was made plausible by the author's skillful characterisation and successful orchestration of the characters' viewpoints. The relationships between characters are narrated by the third-person narrator, who omnisciently takes up each character's perceptions and internal states and reveals their dispositions and motives in life. The analyses through frequent multi-word sequences and an antonymous pair of adjectives show that the narrator takes different but consistent approaches in describing each character's internal states and perceptions.

To collect linguistic material for examination, I used a frequency list of the text of LCL as the primary source. The frequency list and the

aforementioned two keyword lists were made by using the online corpus analysis interface Sketch Engine (Kilgarriff *et al*. 2004).

From the frequency list, the most frequent mental verbs, perception verbs, modal verbs and adjectives were chosen for retrieving the most frequent 2/3-grams. For this purpose, I used the n-gram function of CasualConc, a concordance software for Mac OS X, which was developed by Imao (2011). In addition to these explicit viewpoint marker, two frequent nouns of body parts were also examined in relation to the characters' actions.

These expressions which occur frequently in the text are primarily used for establishing the leading character Connie's viewpoint. These verbs, nouns and adjectives are also applied to present the other main characters' internal states, perceptions and viewpoints although less frequently and vigorously. These characters' inner worlds, compared with Connie's, whose intentions, motives and desires are transparent to the reader, appear to be more vague and distant from the reader. However, after Connie became intimately involved with the gamekeeper, Mellors's viewpoint is more often introduced by similar means to those which were applied to Connie's case. My analysis has revealed how these similar means introduce different characters' viewpoints in different ways. Because these lexical items and phrases are dispersed all over the text, even if they appear more often than other words of the same parts of speech or types, it would not be easy to examine how they work in the text without automated concordance processes. This corpus study shows that particular lexis and phrases permeate the text of LCL and they are consistently applied in representing what the characters see, feel and perceive through their sensory and cognitive capacities. The data of frequencies and contents of these representations also suggest that the uses of these devices can influence characterisation and the degree of empathy of the reader for each character.

As an example, the mental verb which most frequently occurs in the novel is 'know' (345 times). Two/three-word sequences containing this verb which occur more than 10 times were examined in their concordance lines and in the contexts in which the phrases occurred. The combinations of some of the two/three-word sequences and the involved characters (as the subject of the verb or the addressees of the direct speech) show how the characters perceive others and their surrounding worlds or focus on what they desire. Table 1 shows the frequencies of the two/three-word sequences involving 'know' which indicate the characters' knowledge and viewpoints. The verb 'know' is most frequently collocated with 'she' as 'she knew', which occurs 33 times. Out of the 33 cases, in 28 cases the reader is

exposed to what Connie knew: primarily about her own inner states of mind and her understanding of the characters who were close to her. Mellors, the gamekeeper is also collocated with the verb 'knew' 15 times.

	Connie	Mellors	Clifford	Mrs Bolton	others	total
she knew	28	0	0	5	0	33
don't know	6	8	2	8	3	27
he knew	0	15	5	0	4	24
didn't know	9	3	1	0	1	14
I knew	0	8	0	3	1	12
total	43	34	8	16	9	110

Table 1 The frequencies of two/three-word sequences of 'know' and the characters referred to

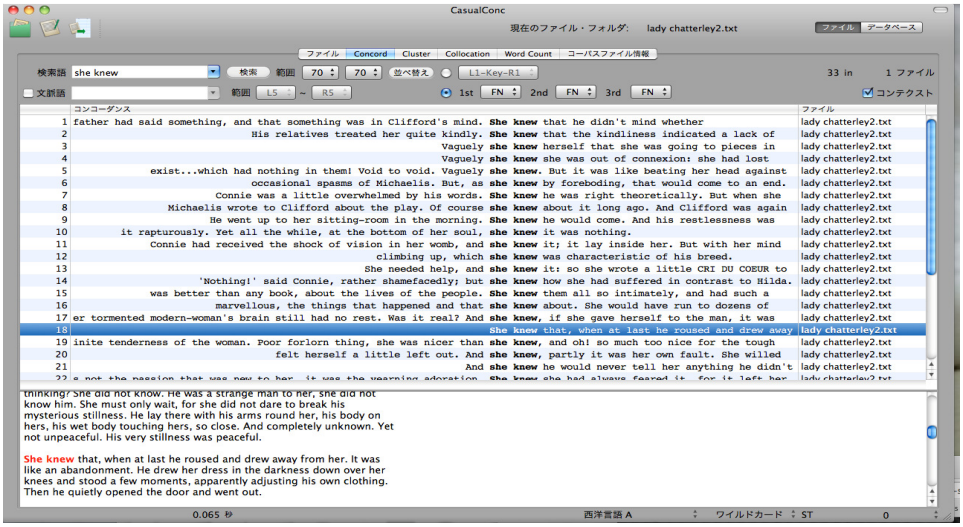


Figure 1 Concordance lines: *she knew*

Connie's inner states of mind are often depicted by the third-person narrator. The 28 concordance lines of 'she knew' are used either to describe her own physical or emotional states that she was aware of or to present her understandings about people around her or events she was involved in. Figure 1 shows the concordance lines of 'she knew'.

Another frequent two-word sequence of 'know' is 'he knew', which occurs 24 times. Out of them, 15 cases refer to what Mellors knew. While by 'she knew' Connie's self-awareness of her inner states or knowledge about the immediate situations which surrounded her were presented, 'he knew' shows Mellors's knowledge about the world in general and assumptions about the world surrounding him.

By retrieving concordance lines using the two/three-gram function, expressions including frequent mental verbs, perception verbs, body-part nouns and a pair of antonymous adjectives can be systematically compared according to each respective character. The analyses show that the leading characters' viewpoints are consistently presented by the repetitive uses of these and other phrases. In particular, Connie's viewpoint is closely coordinated and set parallel with Mellors's. These characters' interrelated viewpoints result in their motives for the romance being convincing and make the structure of the narrative tight.

References

Imao, Y. 2011. Mac OS X no konkodansa CasualConc—kihontekina tsukaikata to

yoreikensaku tsuru to shiteno oyorei [CasualConc, a concordance software for Mac OS X—basic functions and how they can be utilized as a research tool]. *Gaikokugo Kyoiku Media Gakkai Kansai shibu Mesodorogi kenkyu bukai 2011nendo Ronshu* [Journal of Methodology SIG, Kansai Chapter, Japan Association for Language Education and Technology Kansai chapter 2011], 121-178.

Kilgarrif, A., Rychly, P, Smrz, P. & Tugwell D. 2004. "The Sketch Engine".

Proceedings of Euralex, 105-16.

Lawrence, D. H. 1960 [1928]. *Lady Chatterley's Lover*. London: Penguin Books.

A linguistic analysis of 'spin' in health news in English language media

Ersilia Incelli
University of Rome Sapienza
ersilia.incelli@uniroma1.it

1 Introduction

Bridging the communication gap that exists between the scientific community and the public is particularly important in medical research, due to the impact of its findings on the public. The popularization of science tries to span this gap (Calsamiglia and van Dijk 2004). In this process (science/medical) journalists play a crucial role as mediators who transfer and convey medical information and advances in medicine to disseminate 'new' news to their audiences. However, a number of recent studies in the fields of journalism (Goldacre 2011) medicine (Boutron et al. 2012) and linguistics (Fairclough 2006; Suhardja 2009, *interalia),* have pointed at the variable quality of health stories in mainstream media, (particularly those conveying new drugs and new medical procedures), drawing attention to common flaws in health news reporting, such as, lack of attention to the quality of research evidence, exaggerated estimates of the benefits, failure to identify unbiased expert sources. Furthermore, narrative frames often 'distort' the information leading to misinformation and scaremongering, e.g. the 'hype and hope' in stem cell research (see Behnind the Lines NHS report 2011).

2 Research objectives

Therefore, the general aim of the study is to understand the nature of linguistic 'spin' in these reports, and show how 'spin' can be explained by aspects of genre through a comparative analysis. 'Spin' is taken here to involve linguistic notions of sensationlism, bias, language manipulation (Ransohoff and Ransohoff 2001), and from the medical standpoint it is defined as specific reporting strategies (intentional or unintentional) emphasizing the beneficial effect of an experiment or treatment (Boutron et al. 2012). Thus, the overall aim is to investigate science news genre, adopting a genre analysis approach to textual exploration, with a view to comparing three sub-corpora of collected texts: one consisting of medical research papers in which scientists first report their results (e.g. *BMJ, Lancet*), the second consisting of press releases issued from scientific institutions (e.g. pharmaceutical companies, university research laboratories, accessed mainly from the *EurekAlert* database of

science-related press releases), and the third corpus consists of online media texts (namely online newspapers, e.g. *Daily Mail, The Guardian)*. The collected texts were selected according to the most popular health news in the media, e.g. *the flu, heart disease, backache*, etc., and used as case studies.

3 Theoretical and methodological frameworks

The corpora and collected documents reflect the transposition process of scientific results into new contexts or new sources of information. More specifically, the analysis explores the language used to recontextualize and reconstruct the findings of medical research into the context of the 'news story', by identifying the prominent lexico-grammatical and semantic patterns specific to the genre and the pragmatic function of these patterns, which will in turn reveal the most frequent rhetorical 'spin' strategies, recontextualized according to the register and repertoire of the particular genre. All three genre aim to persuade a specific target audience, but each genre has a different communicative purpose with a different audience in mind, according to the level of expertise (Hyland 2010). The transposition process implies different language structures and different moves according to the genre, each occurring with the respective lexical grammatical semantic choices. This also involves an analysis of evaluative language and stance sometimes contributing to bias claims (Hunston and Thompson 2000).

The theoretical underpinnings are obviously in the realm of ESP genre analysis, also referring to Hyland's (2010) 'concepts of proximity', which considers how writers position themselves in relation to others. The study involves background notions of news genre (Galtung and Ruge 1973), news values (Bednarek and Caple 2012), and critical discourse analysis (Fairclough 2006). There is a strong focus on the corpus-assisted approach (Baker 2006; Partington 2010), integrating descriptive qualitative manual analysis with standard corpus linguistic retrieval techniques, such as word frequency lists, concordance and collocation analyses, concgram and ngram retrieval. In fact, the first stage of the analysis focuses on identifying an initial group of keywords, e.g. *new, significant, effective safe, well-tolerated, significant breakthrough, cure, hope, drugs, mice,* which then guided the extraction of reoccurring lexical and collocation patterns, as well as n-grams and phraseolgoical units, e.g. *statistically significant results, more likely to / less likely to, the treatment achieved effective levels of*. The rhetorical statements can then be grouped according to categories of 'spin' strategies, e.g. claiming novelty, claiming validity, making predictions, focusing on secondary outcomes rather than the primary aims of the research, ruling out adverse effects and emphasizing the beneficial effects.

The results so far show that rhetorical spin strategies identified by prominent phraseological patterns which begin in the research paper, are boosted or transposed by the press release, and arrive at the newspaper desk to be further recontextualized, sometimes into exaggeration. What is more, so far most spin appears to originate in the press release (i.e. the institution behind the press release). The journalist then finishes off the 'new' information according to the expected audience interests.

References

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Bednarek, M. and Caple, H. 2012. 'Value added': Language, image and news values.' *Discourse, Context, Media* 1: 103-113.

Behind the Lines, NHS special report. 2011. Available at:http://www.nhs.uk/news/2011/11November/Documents/hope_and_hype_1.0.pdf

Boutron Y.A., Bafeta A., Marroun I., Charles P. et al., 2012. "Misrepresentation of Randomized Controlled Trials in Press Releases and News Coverage: A Cohort Study". *PLoS Med* 9(9): e1001308.

Goldacre, B. 2011. "How far should we trust health reporting?" *The Guardian*, 11th June.

Calsamiglia, H. and van Dijk, T. A. 2004. "Popularization Discourse and Knowledge about the Genome". *Discourse and Society*, 15 (4).

Galtung, J. and Ruge, M. 1973. "Structuring and selecting news". In J. Young and S. Cohen (eds.), *The Manufacture of News: Social Problems, Deviance and the Mass Media* (pp. 62-72). London: Constable.

Fairclough, N. 2006. *Discourse and Social Change*. Cambridge: Polity Press.

Hunston, S. and Thompson, G. 2000. *Evaluation in Text: authorial stance and the construction of discourse*. Oxford: Oxford University Press.

Hyland, K. 2010. "Constructing proximity: Relating to readers in popular and professional science. *Journal of English for Academic Purposes* 9: 116- 127.

Partington, A. 2010. 'Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS) on UK newspapers : an overview of the project'. *Corpora* 5 (2): 83–108.

Ransohoff, D. F. and Ransohoff, R. M. 2001. Sensationalism in the media: When scientists and journalists may be complicit collaborators. *Effective Clinical Practice*, 4: 185-188.

Suhardja, I. 2009. *The Discourse of 'Distortion' and Health and Medical News Reports: A Genre Analysis Perspective.* Ph.D. Thesis, University of Edinburgh.

A phraseological approach to the shift from the *were*-subjunctive to the *was*-subjunctive: Examples of *as it were* and *as it was*

Ai Inoue
National Defense Academy
aiinoue@nda.ac.jp

1 Introduction

It has been generally believed that the *were*-subjunctive used in the phraseological unit[52] *as it were* is strictly prohibited from being substituted with the *was*-subjunctive; however, as examples (1) and (2) show, *as it was* [53] is observed in contemporary English.

(1) MORGAN: Will Justin Bieber have that, do you think? Is it inevitable?
D. OSMOND: He's got it now. He's got it now. You know, that kind of success at that age can really bite you in the shorts, *as it was*, the proverbial shorts.
MORGAN: What would you say to him?
(Corpus of Contemporary American English (COCA), 2011)

(2) The journal had been intended as the perfect Austenesque birthday gift for my vintage-obsessed younger cousin. I'd found it lying alongside a worn copy of Pride and Prejudice in a quirky antiques shop down on South Congress and simply couldn't pass it up, hobnobbing*, as it was,* with greatness.
(COCA, 2012)

As it was in (1) is used to give an example. In (2), it is used to compare the fact that the author found a worn copy of Pride and Prejudice to hobnobbing with greatness.

The purpose of the study is to descriptively show that *as it were* changes into *as it was* from a phraseological perspective. In addition, based on the data collected from corpora, this study minutely explains the actual behaviours of *as it was* and its relationship with *as it were*.

[52] This study defines phraseological units as frequently used combinations consisting of at least two words. Also, the study adopts the most widespread term, 'phraseological units', although various terms such as 'phraseme' and 'recurrent word-combinations' are also used.

[53] *As it was*, which is the focal phraseological unit of this study, of course excludes *as it was* as the *was*-indicative (e.g. *I left it as it was*).

2 Phraseology

Phraseology, the study of phrases, is based on the idea that phrases play essential roles in allowing language activity to proceed smoothly. When reading a newspaper, we can easily find numerous phrases consisting of familiar words. We feel as if we understand their meanings, even if they are not described in dictionaries. Nevertheless, we do not fully understand their meanings. Such phrases are ubiquitous in language, constituting a significant resource for communication. They can also help learners of English as a Foreign Language (EFL) to make their English as fluent as that of native speakers. Most previous studies on phrases have not provided any comprehensive discussion on such phrases. When they are addressed, the discussion usually centres on a single phrase.

The increasing attention on phraseology is due to factors such as the advancement of corpus linguistics, growing interest in lexicography, application of phraseology for language education and advent of language technology applications, such as full text translation, word processing, and text mining. The syntactic rules and lexicons mentioned in existing linguistic theories are less suited than phraseology for explaining the roles that many phraseological units play in contexts involving the acquisition and use of language. Thus, the naturalness of a given language apparently rests on the use of phraseological units. It is obviously possible to generate an infinite number of sentences using the syntactic rules and lexicons explained in linguistic theories. Even if they are grammatically correct and semantically congruent, however, there is no guarantee that sentences generated in this manner will possess the characteristic 'Englishness' reflected in sentences formed by native speakers. Research in English phraseology focuses on identifying the phrases that constitute Englishness and that help language activities to resemble those of native speakers.

The existing phraseological research can be classified into the following two types: one investigates phrases which have long existed, and the other explores newly observed phrases. This study is part of the attempt to re-examine one of existing phrases, *as it were*.

3 Previous research on the *were*-subjunctive

It is said that the *were*-subjunctive tends to be substituted with the *was*-subjunctive (e.g. If I *were/was* rich, I would buy you anything you wanted (Quirk et al. 1985: 168)). According to *Webster's Dictionary of English Usage*, historically, the *was*-subjunctive, instead of the *were*-subjunctive,

began to be used at the end of the 16th century and was frequently used at the end of the 17th century. The dictionary also mentions that the *was*-subjunctive was used for emphasis, but actually, the examples of the *was*-subjunctive appeared in a less formal style.

Schibsbye (1970) explains that the degree of uncertainty greatly influences the choice of either *is*, *was* or *were* and changes depending on *is/was/were*. For example, let us consider the sentence *'If it is/was/were true, I should know it*'. When the *is*-indicative is used, e.g. *if it is true*, it implies that it is obvious that it is true. On the other hand, *if it was true* implies that it is difficult to say whether it is true, and *if it were true* shows that it is not true at all. However, it has been widely acknowledged that the *were*-subjunctive cannot be replaced by the *was*-subjunctive in the case of phraseological units such as *if I were you* and *as it were* (Jespersen 1954; Greenbaum and Whitcut 1988; and Sinclair (ed.) 1992).

4 Change from *if I were you* to *if I was you*

Examining the examples obtained from corpora, we see the interesting phenomenon of the *were*-subjunctive being replaced by the *was*-subjunctive in the case of phraseological units. Instead of *if I were you*, *if I was you* is observed, as elucidated in example (3).

(3) The woman looked at her friend and back to Charlotte. "If I was you, I wouldn't be out walking in this weather unless I had somewhere I had to get to," the woman said.
(COCA, 2012)

Frequencies retrieved through corpora (COCA and the British National Corpus (BNC), as of 30 November and 1 December. 2014) are shown in Table 1.

	if I were you	*if I was you*
COCA	378	59
BNC	152	37

Table 1. Frequencies of *If I Were You* and *If I Was You* in COCA and BNC.

We can see from Table 1 that the frequency of *if I was you* is less than that of *if I were you*. However, it is safe to assume that the understanding that the *were*-subjunctive can be substituted with the *was*-subjunctive is true of *if I were you* regardless of the registers in which it is used.

This phenomenon can be accounted for by the merging of *was* and *were*.

5 From as it were to as it was

Previous research deals with only *as it were*, and no substantive research on *as it was* has been conducted. For example, an English dictionary describes that *as it were* is 'used in order to make what a speaker is saying sound less definite, which means vagueness' and explains that it is used at the middle or end of a sentence working as a sentence modifier (e.g. *Mandela became, as it were, the father of a nation* (*Macmillan English Dictionary* 2nd edition), *If he still refuses we could always apply a little pressure, as it were* (*Cambridge Advanced Learner's Dictionary* 4th edition)). According to the *Oxford English Dictionary (*2nd edition), *as it were* is the abbreviation of *as if it were so*.

Data obtained from corpora reveal that *as it was* is a polysemous phraseological unit. Similar to *as it were*, *as it was* tends to be located at the middle or end of a sentence. Please observe the following examples.

(4) And you're never going to get the politics out of politics, which is where this is and where the outcome, dictated in part*, as it was,* by the political act of holding up a nomination, was manifested. (COCA, 2006)

(5) She treated the animal like a child*, as it was,* and it would only make her defend him more.
(COCA, 1990)

(6) KING: But I mean, you feel the tenets of your church - I don't want to put words in your mouth - you feel the tenets of your church? You are a believer?
Mr. GIBSON: Yeah, yeah, *as it was*.
(COCA, 1990)

In (4), *as it was* is used to give an example of what kinds of acts dictate the outcome. In (5), it is used to compare the animal to a child. In (6), it can be paraphrased to 'as if I was a believer'.
Consequently, the semantic and syntactic features of *as it was* can be summarised in Table 2.

function	*syntactic feature*
give an example	middle of a sentence
compare someone or something to another	middle of a sentence
paraphrase	end of a sentence

Table 2: Semantic and Syntactic Features of *As It Was.*

References

Greenbaum, S. and Whitcut, J. 1988. *Longman guide to English usage*. London: Longman.

Jespersen, O. 1954. *A modern English grammar on historical principles - Part IV - Syntax*. London: Allen &Unwin.

Quirk R., Greenbaum, S., Reech, G. and Svartvik, J. 1985. *A comprehensive grammar of the English language*. London: Longman.

Schibsbye, K. 1970. *A modern English grammar.* Oxford: Oxford University Press.

Sinclair, J.M. (ed.) 1992. *Collins COBUILD English usage for learners*. London: HarperCollins.

Building a Romanian dependency treebank

Elena Irimia
Romanian Institute for Artificial Intelligence, Romanian Academy
elena@racai.ro

Verginica Barbu Mititelu
Romanian Institute for Artificial Intelligence, Romanian Academy
vergi@racai.ro

1 Introduction and state of the art

The growing trend in using syntactical analysis in computer applications dedicated to emulate language-related human faculties (Machine Translation, Summarization, Question Answering, etc.) and the scarcity of resources (annotated electronic corpora) and tools (parsers) covering automatic syntactic analysis for Romanian motivate the work described below. Our purpose is to build a core of a **treebank** for Romanian, comprising 5000 annotated sentences from various functional styles.

We choose the dependency grammar formalism (Tesnière, 1959; Mel′čuk, 1987) considering the characteristics of its structure: *minimal* (each node of the structure is a word in the analysed sentence; there are no artificial nodes in the structure, not even gaps), *ordered* (the order of the nodes reflects the order of the words in the sentence), *multiple branched* (the root of the sentence is the main verb and all its dependents, regardless their number, are attached to it). Moreover, given the relatively free word order of Romanian, the dependency formalism is well suited for syntactic analysis.

Other treebanks were developed for Romanian by:

- Hristea and Popescu (2003): 4042 dependency trees, journalistic genre, short sentences (8,94 words/sentence on average), main clauses only, offering an incomplete description of the language;
- Bick and Greavu (2010): using a rule-based parser[54] with an adapted Italian Constrained Grammar, over 21 million words, not available for download and use in language-based applications;
- Perez (2014): 4500 dependency trees, different functional styles, different historical periods, 37 words/sentence on average. There are intentions to harmonize our treebank and Perez's into a gold-standard treebank for Romanian.

[54] VISL, http://beta.visl.sdu.dk/visl/about/

2 Assembling the corpus

ROMBAC is a Romanian balanced corpus, lemmatised and POS-tagged, developed at RACAI (Ion et al., 2012) and freely available through the METASHARE platform (http://www.meta-share.eu/). It covers four functional language styles (belles-lettres, publicist, official and scientific), distributed in five sections (prose, juridical, medical, academic and journalistic) and we envisage that its syntactic annotation could offer a scale model of the syntactic patterns in the Romanian language.

From this corpus we extracted the 500 most frequent verbs in each sub-section. Naturally, some verbs occur in more than one sub-domain, allowing us the study of their syntactic-semantic behaviour in different linguistic registers. For each selected verb, we automatically recovered from ROMBAC 2 sentences complying with the conditions: 1) each sentence has more than 7 and less than 30 words; 2) each sentence has at least a main verb. Thus, we compiled a corpus of 5000 sentences.

3 Comprehensive dependency relations set for Romanian

Due to Romanian-Spanish typological similarity, our strategy of annotation and correction has been extensively based on the work developed by the IULA[55] team (Arias et al. 2014), that used a Spanish treebank to bootstrap a Catalan treebank. We harmonized our principles and conventions of analysis with theirs and we sought to keep some of the labels they used, aiming to facilitate our manual correction work. We also kept our inventory in line with the strategy used in the Universal Dependency (UD) Project [56] contemplating future multilingual projects and strategies in which we should integrate our efforts. New labels, marked in Table 1 in bold, are justified by language specific syntactical phenomena or by our desire to differentiate between certain relations:

- Pronominal clitics classification: doubling (*dblclitic*), possessive (*posclitic*), reflexive and reciprocal (both as *reflclitic*).
- Differentiation between two co-occurring accusative verb arguments: direct object (identified by the possibility of clitic doubling and of taking the preposition *pe*) and secondary object (*secobj*) (e.g.: *L-au ales pe Ion primar.* "They have elected Ion mayor", where *pe Ion* is the dobj (doubled by *L-*) and *primar* is secobj.)
- Whenever a preposition links the preceding word to the following one (e.g.: *El este teribil de timid.* "He is terribly shy."), we call the relation between the preposition (*de*) and the head (*timid*) *post*.
- When a dependent enters a ternary (both syntactic and semantic) relation, with the verb and with a nominal (the subject or an object), it is assigned to the verb and assigned the label *spe* (supplementary predicative element): *I-am văzut pe copii împreună.* "I have seen the children together.", where *împreună* is *spe* for the verb.
- Conjunctions are treated depending on their type: coordinating ones are attached to the first conjunct by the *cc* relation, while subordinating ones are treated as head of the subordinated clause, whose verb is attached to the conjunction by the *sc* relation: *Vreau să vii*. "I want you to come.", where *să* is *dobj* for the verb *vreau*, while *vii* is *sc* for the conjunction.
- Correlative elements are analysed as correl: *A iubit-o fie pe Maria, fie pe Ana.* "He loved either Maria or Ana.", where the first *fie* is correl for the second conjunction.

4 Annotation and correction

Following closely the procedure described in (Arias et al. 2014), we automatically annotated our corpus using the statistical freely available parser MaltParser [57] (Nivre and Hall, 2005) – with a statistical delexicalised model extracted from Spanish IULA LSP Treebank[58] (Marimon and Bel, 2014) – and manually corrected the resulting trees. Originally, the IULA team has successfully used this approach to boost the creation of a Catalan treebank, motivated by: 1. The typological similarity between Catalan and Spanish; 2. The very good Labelled Attachment Score (LAS, 94%) obtained for the Spanish model when used on Spanish sentences; 3) MaltParser's possibility to construct models controlling different features, e.g. excluding lexical information and using only POS tags. As they are part of the same language family (Romance), we assume the typological similarity between Spanish and Romanian can be exploited to reduce the amount of manual annotation work.

[55] IULA= Institut Universitari de Linguistica Aplicada, Universidad Pompeu Fabra, Barcelona
[56] https://code.google.com/p/uni-dep-tb/
[57] http://www.maltparser.org/
[58] http://www.iula.upf.edu/recurs01_tbk_uk.htm

Romanian	IULA	UD
acl		acl
advcl		advcl
advmod	MOD	advmod
agc	BYAG	agc
amod	SPEC	amod
appos	MOD	appos
aux	AUX	aux
auxpass		auxpass
cc	COORD	cc
compound		compound
conj	CONJ	conj
correl		
dblclitic		
dep	unknown	dep
det	SPEC	det
dislocated		dislocated
dobj	DO	dobj
foreign		foreign
goeswith		goeswith
iobj	IO	iobj
list		list
mark		mark
mwe		mwe
name		name
discourse		discourse
neg	NEG	neg
nmod	MOD	nmod
parataxis		parataxis
passmark	PASSM	
pmod	MOD	
pobj	OBLC	
poss		poss
possclitic		
post		
pred	PRD, ATR	
prep	COMP	case
punct	PUNCT	punct
reflclitic		
remnant		remnant
reparandum		reparandum
root		root
sc		
secobj		
spe		
subj	SUBJ	nsubj, csubj, cubjpass
voc	VOC	vocative
xcomp	OPRD	xcomp

Table 1. Inventories of relations: Romanian, IULA, UD

In order to use the Spanish model, we had to map the tagset[59] we used in the POS-tagging phase for our corpus with the tagset[60] used by IULA (derived from the EAGLES specifications) and also convert our .xml formatted files to the CONLL format used by MaltParser.

Before correction, those IULA labels that could be unambiguously transferred to our dependency label set were automatically mapped accordingly. Labels like SPEC or MOD (with more than one equivalent labels in the Romanian set) were left to be disambiguated by correction. For correction, we used the YeD XML graph editor and some supporting scripts for importing/exporting MaltParser results to/from XML provided by the IULA team.

5 Preliminary evaluations and conclusions

Still at the beginning of our correction work, we present evaluation results for only 100 sentences, part of the journalistic sub-section of our collection of 5000 sentences. We used MaltEval free software, which is an adaptation of the CoNLL evaluation scripts eval.pl and eval07.pl provided by the 2006 and 2007 shared tasks' organizers (Nillson and Nivre, 2008). Traditionally, the syntactic analysis performance is computed in terms of the labelled attachment score (LAS) (number of words with correct heads and labels/number of words), but other measures like LA (number of words with correct labels/number of words), UAS (number of words with correct heads/number of words), etc. are available.

The scores in Table 2 indicate that an important number of manual corrections are imperrative: some error rate is presumed, since we apply a statistic approach with a delexicalised model dedicated to another language; many of the errors (as the correction experience teaches us) are due to the more refined label set we designed, but some are consequences of the different principles of analysis that we and the IULA team applied: e.g. in our approach, an auxiliary verb can never be the head of a sentence. Yet we appreciate the automatic syntactic analysis provided by the IULA Spanish syntactical model is a useful backbone for our correction work (0,715 for either label or head match, AnyRight). According to the Spanish-Catalan experience, after a manual correction of 1000 sentences, a first Romanian module will be trained and we expect an improvement in results comparable to that achieved for Catalan (4% increasing in LAS).

[59] http://nl.ijs.si/ME/V4/msd/html/index.html, MultText East Morphosintactic Specifications for Romanian

[60] http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

Metric	Score
LAS	0.216
LA	0.417
UAS	0.514
AnyRight	0.715

Table 2. Evaluation results for 100 corrected sentences

Acknowledgements

This paper is supported by the Sectorial Operational Programme Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contract number SOP HRD/159/1.5/S/136077".

References

Arias, B., Bel, N., Fomicheva, M., Larrea, I., Lorente, M., Marimon, M., Mila, A., Vivaldi, J. and Padro, M. 2014. "Boosting the creation of a treebank", In *Proceedings of LREC 2014*, Reykjavik, Iceland

Bick J. and Greavu, A. 2010. "A Grammatically Annotated Corpus of Romanian Business Texts", in *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*, Editura Academiei Romane, p. 169-183.

Hristea, F., Popescu, M. 2003. "A Dependency Grammar Approach to Syntactic Analysis with Special Reference to Romanian", in F. Hristea şi M. Popescu (coord.), *Building Awareness in Language Technology*, București, Editura Universității din București, p. 9-16.

Ion, R., Irimia, E., Ștefănescu, D. and Tufiș, D. 2012. "ROMBAC: The Romanian Balanced Annotated Corpus". In *Procedings of LREC 2012* Istanbul, Turkey.

Marimon, M. and Bel, N. 2014. "Dependency structure annotation in the IULA Spanish LSP Treebank". In *Language Resources and Evaluation*. Amsterdam: Springer Netherlands. ISSN 1574-020X

Mel′čuk, I. A. *Dependency syntax : theory and practice*, Albany, State University Press of New York, 1987.

Nilsson, J., and Nivre, J. 2008. "MaltEval: An Evaluation and Visualization Tool for Dependency Parsing", In *Proceedings of LREC 2008*, Marrakesch, Morocco.

Nivre, J. and Hall, J. 2005. "Maltparser: A language-independent system for data-driven dependency parsing", In *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT)*, pages 137-148.

Perez, C.-A. 2014. *Resurse lingvistice pentru prelucrarea limbajului natural*, PhD thesis, "Al. I Cuza" University, Iasi.

Tesnière, L. *Éléments de syntaxe structurale*, Paris, Klincksieck, 1959

Examining Malaysian sports news discourse: A corpus-based study of gendered key words

Habibah Ismail
University of Sydney

hism4614@uni.sydney.edu.au

1 Introduction

This study investigates the representation of male and female athletes in Malaysian sports news discourse in order to determine whether the genders are represented in the media fairly or with bias. Focusing on online news media, the study analyses the daily Malaysian sports news with special emphasis on 'dominant' discourses, which involve instances of objectifying, trivialising, and stereotyping female athletes (Caple 2013). Key words analysis is the main corpus analysis method used, and key words identified are further examined by looking at individual concordance lines. The findings will help to identify either fair or biased gender representations in Malaysian sports news discourse. Since most studies in sports and gender have been on Western media, it will be interesting to see where Malaysian sports news discourse sits in the spectrum of gender representation.

2 Sports News Discourse

Research on Western media has found that female athletes are usually associated with emotionality (more emotion words/images; Jones 2006), passivity (Hardin, Chance, Doss, & Hardin, 2002), immaturity (through infantilisation; Aull & Brown, 2013; McDowell & Schaffner, 2011), and familial roles (Aull and Brown 2013). On the other hand, more equitable gender representation in written and visual texts has also been identified in several studies (King 2007; Vincent et al. 2002), although mainly the data used were news reported during major sports events. While past research on sports news discourse has mostly concerned news coverage of major sports events, day-to-day or daily sports news coverage has not been studied much (Eastman and Billings 2000; Bernstein 2002). Due to this dearth of research, daily sports news reporting will be the focus of this study.

3 The corpus and analyses

The first part of this study discusses the written corpus used to examine representations of male and female athletes in sports discourse in Malaysia. The data were collected from the daily sports news of selected Malaysian online newspapers, resulting in a

3 million word corpus. The corpus consists of hard news and soft news concerning different sports and athletes. The corpus was collected in a span of 6 months, during the final quarter of 2013, and early 2014. This corpus is sub-divided into smaller corpora for the purpose of analysis namely: a female sub-corpus for news written about female athletes, and a male sub-corpus for news written about male athletes.

The second part of this study involves an analysis of the key words. The key word list was generated when the female sub-corpus was compared against the male sub-corpus. The findings show that there is a difference between women's sports reports and men's sports reports in terms of different salient words used. Examination of the key words also reveals a difference in the focus of respective news reports: women's sports reports revolve around individual athletes, while men's sports reports focus on team sports. Furthermore, the key words show the prominence of a few star female athletes in the female sub-corpus. This finding concurs with previous studies by Jones (2006) and Markula (2009) who reported that a few female 'star' athletes contribute to the inflation of female coverage frequency.

In the analysis of gender related key words, words that appear as key words in their order of keyness strength, are: pronouns (*she/her*), and gendered nouns (*woman/women/women's*, *girls*, *sister* and *female*). The gendered nouns appeared as key words but not their male counterparts (i.e. *man/men/men's, boys, brother* and *male*). For the sake of comparison, the counterparts to these key words were also analysed. Thus, gendered words that were identified and examined include gendered pronouns (*she, her, he, and his*), the lemmas MAN and WOMAN, GIRL and BOY, FEMALE and MALE, and other related key words.

The third component of this study comprises in depth evaluation on individual concordance lines based on the keywords. This reveals language patterns that are distinct to women's sports reporting. For example, descriptions of female athletes may involve emphasis on physical evaluations. On the other hand, descriptions of male athletes are mostly related to the games and their professional performances. This analysis also demonstrates how language patterns used to describe women's sports are less objective and factual than men's sports. Examination of these key words sheds light on the kinds of words and language patterns that are prevalent in Malaysian sports news discourse: For example a language pattern was identified where the word *men's* is always positioned before *women's* in almost all syntactic sequences. This type of syntactic bias where male is always in the leading position is referred to as male firstness. In the case of binomial pairs, it was argued that people may regard that the entity in the leading position of the pair is the preferred norm out of the two (Baker 2014).

In the examination of Malaysian Sports news discourse, it was found that there are differences in terms of how male and female bodies are articulated in the news reports especially on the different words used to narrate their news stories. Overall, the findings help to identify words and language patterns that insinuate fair or biased gender representation and contribute to the investigation of gender bias in sports news discourse.

References

Aull, L. L., & Brown, D. W. 2013. Fighting Words : A Corpus Anaysis of Gender Representations in Sports Reportage. *Corpora* 8(1): 27–52.

Baker, P. 2014. *Using Corpora to Analyze Gender*. London: Bloomsbury Publishing.

Bernstein, A. 2002. Is It Time for a Victory Lap?: Changes in the Media Coverage of Women in Sport. *International Review for the Sociology of Sport* 37(3-4): 415–428.

Caple, H. 2013. Competing for Coverage: Exploring Emerging Discourses on Female Athletes in the Australian Print Media. *English Text Construction* 6(2): 271–294.

Eastman, S. T., & Billings, A. C. 2000. Sportscasting and Sports Reporting: The Power of Gender Bias. *Journal of Sport and Social Issues* 24(2): 192–213.

Hardin, M., Chance, J., Doss, J. E., & Hardin, B. 2002. Olympic Photo Coverage Fair to Female Athletes. *Newspaper Research Journal* 23(2,3): 64–78.

Jones, D. 2006. The Representation of Female Athletes in Online Images of Successive Olympic Games. *Pacific Journalism Review* 12(1): 108–129.

King, C. 2007. Media Portrayals of Male and Female Athletes: A Text and Picture Analysis of British National Newspaper Coverage of the Olympic Games Since 1984. *International Review for the Sociology of Sport* 42: 187–199.

Markula, P. 2009. Introduction. In P. Markula (Ed.), *Olympic Women and the Media: International Perspectives* (pp. 1–29). Basingstoke: Palgrave MacMillan.

McDowell, J., & Schaffner, S. 2011. Football, It's a Man's Game: Insult and Gendered Discourse in The Gender Bowl. *Discourse & Society* 22(5): 547–564.

Vincent, J., Imwold, C., Masemann, V., & Johnson, J. T. 2002. A comparison of selected "Serious" and "Popular" British, Canadian, and United States newspaper coverage of female and male athletes competing in the centennial olympic games: Did

female athletes receive equitable coverage in the "Games of the Women"? *International Review for the Sociology of Sport* 37(3-4): 319–335.

Doing well by talking good? Corpus Linguistic Analysis of Corporate Social Responsibility (CSR)

Sylvia Jaworska
Reading University
s.jaworska@reading.ac.uk

Anupam Nanda
Reading University
a.nanda@reading.ac.uk

1 Introduction

Given the growing awareness of shrinking resources, pressure is mounting on businesses to relocate some of their profits back to society and to increase activities generating social good. These are normally subsumed under the banner of Corporate Social Responsibility (CSR) and centre around issues of sustainable future, employees' wellbeing and community engagement. The growing body of business literature concerned with CSR points to its wide-ranging impacts including an 'enhanced' image and reputation as well as a positive effect on financial performance (e.g. Roberts and Dowling 2002, Porter and Kramer 2006). However, each study proposes a different 'take' or measure of CSR and hence, findings are often conflicting or difficult to compare across businesses and sectors (see Griffin and Mahon 1997). This lies partially in the difficulty to empirically capture the nature of CSR activities, many of which are mediated textually in corporate disclosures known as the CSR reports.

Examining corporate disclosures is highly relevant because they are the most visible documents describing organisations' actions and goals in relation to its stakeholders and society. There is evidence suggesting that the publication of corporate disclosures and their specific linguistic properties have a tangible impact on company's performance in that they directly influence market responses and investors' decision-making (e.g. Henry 2008, Li 2008). Despite the growing importance of CSR and the impact of corporate disclosures, there has been little research that examined the language of CSR reports. The few studies that exist are limited in scope; they are often concerned with one country (e.g. Lischinsky 2011), based on a small amount of reports (e.g. Tengblad and Ohlsson 2010) or examine broad constructs such as optimism or certainty (e.g. Cho et al. 2010). Most of the studies use the methodology of Content Analysis which has been criticised for its reliance on subjective semantic coding of data determined *a priori*.

2 Research aims and methodology

This aim of this study is to examine the relationship between the mediated textual representations of CSR

and their impact on company's performance. In contrast to previous small-scale research, it is based on a large corpus of CSR reports produced by 20 major oil companies between 1998 and 2013 (corpus size: 14,915,714 tokens). This sector was chosen because of its direct involvement in environmental issues (often disasters) and the resulting public criticism. The main questions which this research addresses are:

- Q1: What are the key messages and topics communicated in the CSR reports?
- Q2: How do they change over time and in response to significant events (e.g. financial crisis)?
- Q3: Is there a relationship between the identified CSR topics and other performance indicators of the studied companies?

Whereas previous research on CSR reports used mainly the methodology of Content Analysis, we adopt the tools and methods of Corpus and Computational Linguistics that are increasingly used to study large amounts of textual data in Social Sciences (e.g. Lischinsky 2014, Riddel 2014) and allow for semantic categories to emerge from the data. Keyword analysis and topic modelling are performed on the data to identify the key messages of CSR reports and their changes over time. These are subsequently correlated with financial data to test whether there is a relationship between the CSR topics and companies' performance. This study is also an example demonstrating how corpus and computational tools and methods can be effectively used to increase our understanding of issues pertaining to business, economy and society.

References

Cho, H., Roberts, R. and Pattens, D. 2010. "The language of US corporate environmental disclosure", *Accounting, Organizations and Society* 35 (4): 431-443.

Griffin, J. and Mahon, J. 1997. "The Corporate Social Performance and Corporate Financial Performance Debate: Twenty-Five Years of Incomparable Research", *Business and Society* 36 (1): 5-31.

Henry, E. 2008. "Are investors influenced by how earnings press releases are written?", *Journal of Business Communication* 45 (4): 363-407.

Li, F. 2008. "Annual report readability, current earnings, and earnings persistence", *Journal of Accounting and Economics* 45: 221-247.

Lischinsky, A. 2011. "The discursive construction of a responsible corporate self". In A.E. Sjölander and J. Gunnarson Payne (eds.) *Tracking discourses: Politics, identity and social change*. Lund: Nordic Academic Press: 257-285.

Lischinsky, A. 2014. "Tracking Argentine presidential discourse (2007-2014): a computational approach". Presented at *CADAAD 2014*, Budapest, Hungary.

Porter, M. and Kramer, M. 2006. "Strategy & society: The link between competitive advantage and corporate social responsibility", *Harvard Business Review* 84 (12): 78-92.

Riddell, A. 2014. "How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models". In M. Erlin and L. Tatlock (eds.) *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, 91-114.

Roberts, P.W. and Dowling GR. 2002. "Corporate reputation and sustained superior financial performance", *Strategic Management Journal* 23 (12): 1077-1093.

Tengblad, S. and Ohlsson, C. 2012. "The Framing of Corporate Social Responsibility and the Globalization of National Business Systems: A Longitudinal Case Study", *Journal of Business Ethics* 93 (4): 653-669.

Representations of Multilingualism in Public Discourse in Britain: combining corpus approaches with an attitude survey

Sylvia Jaworska
Reading University
s.jaworska@reading.ac.uk

Christiana Themistocleous
Reading University
c.themistocleous@reading.ac.uk

1 Introduction

Since the publication of Hardt-Mautner's (1995) corpus-based work on the representations of Europe in the British press, corpus tools and methods have been increasingly used to study media constructions of social groups and phenomena (e.g. Baker and McEnery's 2005, Gabrielatos and Baker 2008, Baker et al. 2013). By combining quantitative corpus techniques with procedures typical for qualitative discourse studies, this research has been invaluable in revealing persistent and also more nuanced patterns of representations disseminated in the media and gradually influencing public opinion.

There is no doubt that media, especially national media, play a powerful role in influencing opinions and issues surrounding language(s) are no exception (Kelly-Holmes 2012). Without questioning this impact, this study attempts to examine more closely the link between the textual media representations and the popular ways of thinking by taking as an example the representations and attitudes towards bi- and multilingualism.

Studying representations of multilingualism in British public discourse is an endeavor of high social relevance. Britain is one of the most linguistically diverse countries in Europe and this diversity is celebrated (Milani et al. 2011). At the same time, the knowledge of languages is considered problematic and sometimes iconically associated with negative events or undesirable forms of behavior (Blackledge 2004). Research concerned with the thematising of multilingualism has shown that media are vehicles of such ambiguous representations in that they tend to reduce the complexity of multilingual practices to a few essentialist images or myths (Ensslin and Johnson 2006, Kelly-Holmes and Milani 2011, Lanvers and Coleman 2013).

2 Research aims

There is already a considerable body of research concerned with the media thematising of multilingualism. However, most of this work examines representations of selected linguistic varieties and with exception of Vessey (2013) and Ensslin and Johnson (2006), it is based on small samples. Equally, sociolinguistic work interested in the attitudes towards multilingualism focuses on specific varieties and is mostly concerned with learners' or parental attitudes.

The focus of this study is not on a particular language or variety, but on multilingualism as a linguistic and social phenomenon. It follows two aims. Firstly, we are interested in the discourses about bi- and multilingualism disseminated in British national newspapers and how they have changed over time. Secondly, we examine the extent to which the media representations are shared and/ or refuted in the views of general public.

3 Research methodology

To accomplish the first aim, we adopted the methodology of Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS) (Partington 2010). To this end, a large corpus of articles from the major British national newspapers discussing bi- and multilingualism MinD (**M**ultilingualism **in** Public **D**iscourse) and published since 1990 was compiled. The articles were downloaded from Nexis UK. The search terms were *multilingual** and *bilingual**. To ensure that bi- and multilingualism were topical and not mentioned in passing, only articles in which these terms occurred 3 times or more were included in the corpus. In order to investigate the diachronic aspect, the corpus was divided into 3 subcorpora each including articles from a different decade (see Table 1). To identify salient discourses surrounding bi- and multilingualism in each decade, keywords were retrieved and the first 100 grouped into semantic categories. Selected keywords were examined via concordances.

Corpus	Tokens	Articles
MinD1 (1990 – 1999)	204,677	195
MinD2 (2000 – 2009)	437,006	302
MinD3 (2010 – 2014)	209,841	167
Total	851,524	664

Table 1: Corpus Size

To address the second aim, an online survey was created and distributed to people living in a large urban city in the South of England. The survey was divided into three parts including: 1) questions regarding the age, gender and the linguistic background of the participants, 2) a series of positive, negative and neutral statements regarding multilingualism to be rated on a Likert-scale, 3) an open-ended question asking participants to express their personal views towards multilingualism and 4)

5 scenarios presenting bilingual individuals in different roles, with different backgrounds and language abilities, to be rated as bilingual also on a Likert-scale. The statements as well as the scenarios were fed by results that emerged from the corpus analysis. 211 participants responded to the survey, of which the majority were female (70.6%). The average age was 31 (ranged from 16 to 78) and 52.1% spoke more than one language (47.9% identified as monolingual). The results were analysed by using the SPSS package.

4 Results

The keyword analysis points to similarities, but also thematic shifts in the discourse about bi- and multilingualism in the last three decades. Generally, bi- and multilingualism are discussed in the context of schooling, which suggests that both are seen predominately as educational 'products'. This is also supported by the prominence of 'children' and 'pupils' that are consistently identified as the *key* social actors across the three decades. Also, bi- and multilingualism are consistently linked with prestigious linguistic varieties useful worldwide such as English, French and Spanish. Interestingly and with exception of Polish, languages spoken by the minorities in the UK do not belong to the 'strongest' keywords.

Semantic Category	Examples of keywords
Education	schools, education, school, teaching, learn, learning …
Languages/ language varieties	English, French, Welsh, Gaelic, Spanish, German, Italian…
Social actors	children, pupils, teachers, students, Canadians, bilinguals …
Countries/regions	Quebec, Canada, Wales, France, European, Britain
Linguistic terms	bilingual, language, languages, multilingual, bilingualism …
Evaluation	foreign, fluent, ethnic, fluently
Communication	speak, speaking, says
Cities	London, Bangor
Medical/bodily terms	dyslexia, dyslexic, deaf
Others	Internet, KGB

Table 2: Keywords in MinD1 (1990-1999)

There are also a number of differences. For example, whereas in the 1990s and 2000s there seemed to be a stronger focus on regional languages in the UK (Welsh and Gaelic) and abroad (Irish) (see Table 2 and 3), these are not the strongest keywords in the current decade (see Table 4). Other differences occur in the category of social actors. It is interesting to note that the item 'immigrants' appears in the last two decades (15.3 per 100,000 in MinD2 and 25.4 per 100,000 in MinD3) indicating that increasingly bi- and multilingualism are linked with immigration (see Table 3 and 4). Examining collocations of the lemma 'immigrant' confirms the negative semantic prosody of immigration revealed in previous research (e.g. Gabrielatos and Baker 2008). In the context of bi- and multilingualism, immigrants are too mostly associated with criminality ('illegal') and large numbers ('influx'). They are also 'young' and 'poor', come predominantly from Africa and Eastern Europe and do not speak English, which is seen as a burden. A further difference concerns the use of medical terms that crop up in the current decade. The collocational profiles of the items 'Alzheimer's' and 'dementia' show that some sources (mostly the tabloid press) tend to portray bilingualism as a preventative measure against this brain disorder supporting the myth that bilingualism can 'cure' dementia. In contrast, broadsheets present a more cautious picture.

Semantic Category	Examples of keywords
Education	school, learning, schools, learn, primary, education
Languages/ language varieties	English, French, Welsh, Gaelic, Spanish, Catalan, Irish
Social actors	children, pupils, speakers, people, graduates, parents, immigrants
Countries/regions	EU, Wales, UK, France
Linguistic terms	language, languages, bilingual, multilingual, bilingualism
Evaluation	foreign, fluent, native, cultural
Communication	speak, says, speaking, translation, spoken
Cities	London, Beijing
Others	online, website, signs

Table 3: Keywords in MinD2 (2000-2009)

Semantic Category	Examples of keywords
Education	school, schools, learning, primary (school),
Languages/ language varieties	English, French, Spanish, Mandarin, German, Flemish,
Social actors	children, pupils, speakers, immigrants, Bialystok
Countries/regions	EU, Malta, UK, Belgium
Linguistic terms	language, languages, bilingual, bilingualism
Evaluation	foreign, fluent,
Communication	speaking, speak, says
Cities	Brussels, Manchester
Medical/bodily terms	Alzheimer's, dementia, brain, cognitive

Table 4: Keywords in MinD3 (2010-2014)

In summary, the keyword analysis has demonstrated a number of constant but also shifting representations of multilingualism. Currently, there seem to be two parallel evaluations. While bilingualism associated with prestigious varieties ('elite' bilingualism) is overall positively valued, bilingualism linked with community languages is increasingly associated with immigration and seen as a 'burden'. The analysis also reveals that some sources tend to reinforce certain misconception e.g.

bilingualism can prevent dementia.

The results from the attitude survey confirm and refute some of the media representations. Overall, the attitudes towards multilingualism were positive with respondents highlighting benefits such as better job opportunities and cultural diversity, though some negativity was expressed too with monolingual speakers being more likely to mention negative aspects e.g. problems with social cohesion, a perceived lack of willingness on the part of 'foreigners' to learn English. Also, the majority of the respondents believed that widely spoken world languages are more useful and that being multilingual meant a high level of fluency in all languages. The endorsement for 'elite' bilingualism was confirmed by the rating of the scenarios. Individuals who had a language qualification and in addition to English, spoke a prestigious variety (French) were more likely to be rated as bilingual than those who spoke a community language (Polish). Conversely, most participants disagreed with the view that bilingualism can prevent dementia demonstrating that this myth is not shared by general public.

References

Baker, P. and McEnery, T. 2005. "A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts". *Journal of Language and Politics* 4 (2): 97-226.

Baker, P., Gabrielatos, C. and McEnery, T. 2013. *Discourse analysis and media attitudes*. Cambridge: Cambridge University Press.

Blackledge, A. 2004. "Constructions of identity in political discourse in multilingual Britian". In A. Pavlenko and A. Blackledge (eds.) *Negotiations of Identity in Multilingual Contexts*. Clevedon: Multilingual Matters, 68-92.

Ensslin, A. and Johnson, S. 2006. "Language in the news: investigating representations of 'Englishness' using WordSmith Tools". *Corpora* 1 (2): 153-185.

Gabrielatos, C. and Baker, P. 2008. "Fleeing, Sneaking, Flooding: A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the UK Press, 1996-20052. *Journal of English Linguistics* 36: 5-38.

Hardt-Mautner, G. 1995. "Only connect: critical discourse analysis and corpus linguistic". *UCREL Technical Paper* 6. Lancaster: University of Lancaster.

Kelly-Holmes, H. 2012. Multilingualism in the Media. In: M. Martin-Jones, A. Blackledge and A. Creese (eds.) *Routledge Handbook on Multilingualism.* London and New York: Routledge, 333-346.

Kelly-Holmes, H. and Milani, T. 2011. "Thematising multilingualism in the media". *Journal of Language and Politics* 10 (4): 467-489.

Lanvers, U. and Coleman, J. 2013. "The UK language learning crisis in the public media: a critical analysis". *Language Learning Journal*, published online 11 Oct 2013.

Milani, T. M., Davies, B. and Turner, W. 2011. "Unity in disunity: Centrifugal and centripetal forces of nationalism on the BBC Voices website". *Journal of Language and Politics* 10 (4): 587-613.

Vessey, R. 2013. "Challenges in cross-linguistic corpus-assisted discourse studies". *Corpora* 8 (1): 1-26.

"Can you give me a few pointers?" Helping learners notice and understand tendencies of words and phrases to occur in specific kinds of environment.

Stephen Jeaco
Xi'an Jiaotong-Liverpool University
smjeaco@liv.ac.uk

Concordancers have great potential as a language learning tool. Besides providing a means for drawing out multitudes of examples showing how words and phrases can be used, Data Driven Learning can provide opportunities to explore and uncover patterns in language use which may not be available from other resources. Bernardini (2004) suggests that with learners in the role of "traveller", concordancing tasks can be used to meet a variety of language teaching goals. Concordancers can be used to show differences between similar words, through searches for pairs of words provided by the learners (Johns 1991) and searches for pairs of synonyms (Tsui 2004; Kaltenböck and Mehlmauer-Larcher 2005). However, while concordancing can be rewarding, feedback from students has also shown that the discovery process can be both difficult and time-consuming (Yeh et al. 2007). In this digital age, university students tend to have less fear of making use of similar resources such as search engines, but introducing concordancers in the classroom or for independent study is still very challenging. One key challenge is helping new users to understand what the nearby context of a node word can show about how a word is typically used. Another key challenge is helping students to develop skills to know what kinds of pattern to look for and how to weigh the evidence. Students of English for Academic Purposes (EAP) are in particular need of more resources so that they can make appropriate choices in their own written output. It has been argued that concordancing software needs to be improved to make it more user-friendly and more suitable for language learners (Krishnamurthy and Kosem 2007). The theory of Lexical Priming (Hoey 2005) provides further challenges to a software developer through its argument that beyond attention to meaning and collocation in order to use language in a natural sounding way, many aspects of the typical environments of words are important. This paper introduces several features of a new software tool which has been designed for self-tutoring and teaching for students of EAP, particularly at the intermediate and advanced levels. It has been designed to help such language learners notice patterns in the typical environments in which words and phrases are used. This is achieved through the design of an extended concordance line display and through "hot" icons which are used to indicate strong tendencies that can be further explored through graph data or concordance line filtering.

The KWIC display available in most concordancers has some important advantages including the number of results that can be viewed together (Mair 2002), the way in which it helps users focus on the "central" and "typical" (Hunston 2002), and the "snapshot" it can provide of how lexis is usually used (Johns 2002). However, it has also been recognised that longer contexts may be needed in order for some kinds of information to be revealed (Sinclair 1991; Hunston 2002). By giving users an alternative view of concordance lines in the form of "cards", this new software makes it easier to see how words and phrases are typically used in terms of textual colligation, with headings and paragraph breaks clearly shown and the position of the node in the sentence clearly evident.

The second feature of the software relates to the display of "hot" icons representing tendencies for use in certain kinds of environment. These appear based on an extension of the key word technique. Hoey and O'Donnell (2008)} and O'Donnell et al. (2012) applied the key word technique in order to measure tendencies of words to occur in text or paragraph initial position by treating the first sentences of texts and paragraphs as a study corpus and the sentences from the remainder of these texts as a reference corpus. The results of the latter study demonstrated that one in forty individual words had a tendency to occur in specific positions, and this provides good evidence that this is something worth researching further. If the starting point, however, is a particular word or phrase which a language learner wants to explore to see whether or not it has such a tendency, it is clear that in roughly thirty-nine out of forty cases the results are likely to be disappointingly negative. Also very few users of standard concordancing software in a language learning setting would have the skills or motivation required to work through the process of dividing sentences in a corpus according to text position, and then of performing key word analysis and interpreting the results themselves. Nevertheless, the key word approach could be applied in order to measure tendencies for a range of linguistic features including textual colligation, and language learners may find information about these tendencies helpful. When discussing concordance software more generally, Cobb (1999) argues that for language learning, software is needed that does not assume detailed linguistic knowledge or assume that the users will be curious enough to explore. The new

software tool aims to avoid these two assumptions.

Fourteen kinds of environment are measuring through the corpus pre-processing scripts, including several measures of textual colligation as well as measures which may come under the headings of colligation or collocation and are related to aspects of grammar and usage with which language learners often struggle in their own writing. Only the icons representing those tendencies which achieve a certain level of statistical significance are displayed prominently on a dock at the bottom of the screen. Clicking on these icons takes the user to a graph display showing the proportion of concordance lines which have this feature and indicators of the "expected" proportions based on the corpus overall. From this display, it is also possible to filter the concordance lines according to whether or not they occur in specific kinds of environment. The filter can be used to compare lines inside and outside the environments as a way of helping users understand what is being measured, and as a way of opening up the potential for these different environments to demonstrate different uses of a word or phrase.

The software has been tested with a range of corpora, including several corpora of academic texts. The dock typically shows 3 to 5 icons for any node word, although in some cases there may be very many and in other cases none at all. As would be expected words and phrases often have different tendencies across different text types.

The project was designed to enable teachers and students to explore various features of the theory of Lexical Priming without needing to teach the theory explicitly. It would not be desirable to offer students a complicated exposition of Lexical Priming with all the technical and linguistic background knowledge which that would require. The software is designed, however, to encourage exploration of some of the features identified in this theory and to make it possible to see tendencies of words and phrases which are not usually apparent in either dictionary examples or the output from other concordancing software.

References

Bernardini, S. (2004). "Corpora in the classroom: An overview and some reflections on future developments". In J. M. Sinclair *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins**:** 15-36.

Cobb, T. (1999). "Giving learners something to do with concordance output". *ITMELT '99 Conference.* Hong Kong.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language.* London, Routledge.

Hoey, M. and M. B. O'Donnell (2008). "Lexicography, grammar, and textual position." *International Journal of Lexicography* **21**(3): 293-293.

Hunston, S. (2002). *Corpora in Applied Linguistics.* Cambridge, Cambridge University Press.

Johns, T. (1991). "Should you be persuaded: Two samples of data-driven learning materials". In T. Johns and P. King *Classroom Concordancing*. Birmingham: Centre for English Language Studies, University of Birmingham. **4:** 1-13.

Johns, T. (2002). "Data-driven Learning: The perpetual change". In B. Kettemann, G. Marko and T. McEnery *Teaching and Learning by Doing Corpus Analysis*. Amsterdam: Rodopi**:** 107-117.

Kaltenböck, G. and B. Mehlmauer-Larcher (2005). "Computer corpora and the language classroom: On the potential and limitations of computer corpora in language teaching." *ReCALL* **17**(01): 65-84.

Krishnamurthy, R. and I. Kosem (2007). "Issues in creating a corpus for EAP pedagogy and research." *Journal of English for Academic Purposes* **6**(4): 356-373.

Mair, C. (2002). "Empowering non-native speakers: the hidden surplus value of corpora in Continental English departments". In B. Kettemann, G. Marko and T. McEnery *Teaching and Learning by Doing Corpus Analysis*. Amsterdam: Rodopi**:** 119-130.

O'Donnell, M. B., M. Scott, et al. (2012). "Exploring text-initial words, clusters and concgrams in a newspaper corpus." *Corpus Linguistics and Linguistic Theory* **8**(1): 73-101.

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation.* Oxford, Oxford University Press.

Tsui, A. B. M. (2004). "What teachers have always wanted to know - and how corpora can help". In J. M. Sinclair *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins**:** 39-61.

Yeh, Y., H.-C. Liou, et al. (2007). "Online synonym materials and concordancing for EFL college writing." *Computer Assisted Language Learning* **20**(2): 131-152.

Researching small and specialised corpora in the age of big data (*panel*)

Alison Johnson (*chair*)
University of Leeds
a.j.johnson@leeds.ac.uk
(panel speakers to be confirmed)

1 Introduction

In the always-on digital world of Big Data and billion word corpora, there is, we argue, still a place for small and specialised corpora (Cameron and Deignan 2003; Flowerdew 2004). Advantages of these corpora are that they allow the researcher to ask domain-specific questions, focus on particular discourse features in those domains, and research significant current world events. This panel features papers that encompass and exploit these advantages.

There are four papers in this panel, two of which focus our attention on corpora collected in response to significant world events – The Arab Spring narrative and the *Je suis Charlie* story – in the news and in social media, and two which deal with small corpora created from *The Proceedings of The Old Bailey (POB)* (Hitchcock et al 2012), a 120 million online corpus. The panel, therefore, examines representations of the past and the present in relation to issues of social and legal importance.

2 The small and specialised corpora in this panel

The two small and specialised corpora created from the *POB* are a corpus of around 150 18th century rape trials, totalling around 380,000 words, and a corpus of around 1,000 19th century trials, which feature the prosecution and defence advocacy of a single barrister, Mr Horry of Gray's Inn, totalling around 1 million words.

The two corpora that are centred on the present day are first of all: The Arab Spring Corpus, a corpus of news texts in both English and Arabic totalling 5.9 million words and 5.6 million words respectively, and made up of a total of more than 15,000 texts, both news and editorial. It is time-bound, starting before the first use of the words *Arab Spring* on 15 June 2010 and ending on 31 August 2013, after the Turkish uprisings. Secondly, a corpus of the recently-breaking *Je suis Charlie* story is being collected. The *Je suis Charlie* corpus contains news and social media texts and is also time bound, starting on 7 January 2015 and ongoing at the time of writing. This corpus is being worked on by a large team of staff and students and the results obtained from this teamwork will be shared by a representative from the group.

3 The research questions and methodologies

The *POB* feature historical trial discourse and the research question posed in relation to both corpora is the same: What are the legal pragmatic functions of reverse polarity questions? Questions play an important controlling and constraining role in any trial and lawyers use different questions for a range of legal functions and in order to bring about important legal goals: supporting a prosecution or undermining that prosecution in a defence. The role of reverse polarity questions, such as: Why did you not cry out? Why did you not scream when he let you go? Why did you not tell her [the mother] then? in the rape trials, control the complainant's narrative in ways that are detrimental to the prosecution and advantageous to the defendant, casting the rape narrative in stereotypical and narrow ways. There are over 100 such questions in the rape trials (which are recorded in question and answer format) and, though they do not feature frequently in individual trials, looking across the corpus of trials enables us to focus on this single coercive question type, accounting for its use through the analysis of multiple examples. Similarly, the Horry corpus, collected using the Old Bailey website API tool, contains around 100 examples of this question type. Although the 19th century trials are not recorded in question and answer format, with only a small number of questions recorded and the record comprising mostly the answers to questions, some of the questions **are** recorded. The fact that they have been recorded makes them interactionally salient, in that the court reporter and/or the editor of the OBP publication deemed the questions important to include. So, in this corpus of around 1,000 trials, in which Mr Horry acts as either prosecution or defence barrister, the use of these questions can still be studied. Most of the reverse polarity questions appear in cross-examination discourse, so analysis focuses on the function of reverse polarity questions in this activity, such as in Figure 1, sorted to show Mr Horry asking the question in cross-examination.

N	Concordance
14	." Cross-examined byMR. HORRY. Did you not tell anybody about it? Not till
15	. Cross-examined byMR. HORRY. Have you not been something else here
16	. Cross-examined byMR. HORRY. Did you not find some flour bags also? No.
17	. Cross-examined byMR. HORRY. Did you not form your belief when you

Figure 1. 4 of 99 lines of a concordance for you not, showing reverse polarity questions in cross-examination.

The research questions asked in relation to the contemporary corpora centre on the identification and representation of key social actors. In the Arab Spring Corpus the question is: Who is being reported?, asking who the main news actors are in

the English and Arabic news media before and after the emergence of the Arab Spring. Qualitative and quantitative approaches are used to investigate the similarities and differences in the representations of key social actors in the two text types in this corpus: news and editorial. In the Je suis Charlie corpus the research question is: How are the key political, social and religious actors represented in the corpus? *Wordsmith Tools* (Scott 2011) and *CFL Lexical Feature Marker* (Woolls 2012) are the corpus tools employed alongside qualitative (critical) discourse analysis and a CADS approach (Partington 2006).

4 Findings and results

In the historical corpora we have identified the controlling and coercive work that is done by reverse polarity questions, but also the key placement of these questions in cross-examination discourse, making them a particular resource of witness narrative destruction. Drawing on a discourse-historical perspective we show how current rape myths are grounded in a long social history and persistently present in 18th century court proceedings at the Old Bailey. In the 19th century, in the cross-examination advocacy of a single lawyer, we see how these questions form an important resource in one lawyer's strategic toolkit and show the particular ways that he uses these questions to open and continue his cross-examination.

In the contemporary period, drawing on different discursive strategies of sociolinguistic representation in discourse such Fairclough's (1995, 2001) notion of foregrounding and backgrounding, van Leeuwen's (2008) notion of a 'social actor network' (exclusion/inclusion strategies) as well as van Dijk's (1998) ideological square, the results of the contemporary studies are as follows. In the Je suis Charlie corpus, work is just beginning (at the time of writing), but in the Arab Spring Corpus we have identified significant differences in the representations (negative and positive) of the key news actors, be it of elite and powerful or ordinary people, in terms of lexical choice, labels and stereotypes, as well as the topics with which they were associated. This indicates that many of the Arab Spring news stories are politically, socially and ideologically polarized, and that the mass media in general, and the news media in particular play a significant role in constructing social reality as well as in determining what to include and/or exclude and how social/news actors and their activities are represented.

References

Cameron, L., Deignan, A. 2003. "Combining large and small corpora to investigate tuning devices around metaphor in spoken discourse". *Metaphor and Symbol* 18 (3): 149-160.

van Dijk, T. A. 1998. Opinions and Ideologies in the Press. In: A. Bell, A. and P. Garrett (eds.) *Approaches to Media Discourse.* Oxford: Blackwell, 21-63.

Fairclough, N. 2001. *Language and Power*. 2nd edn. London and New York, Routledge.

Fairclough, N. 1995. *Media discourse*. London, Edward Arnold.

Flowerdew, L. 2004. "The argument for using English specialized corpora to understand academic and professional language". In Connor, U. and Upton, T. A. (eds) *Discourse in the Professions.* Amsterdam: John Benjamins, 11-33.

Hitchcock, T., Shoemaker, R., Emsley, C., Howard, S. and McLaughlin, J. 2012. *The Old Bailey Proceedings Online, 1674-191.* www.oldbaileyonline.org, version 7.0, 24 March 2012 [accessed 14 January 2013].

van Leeuwen, T. 2008. *Discourse and Practice: New Tools for Critical Discourse Analysis.* Oxford: Oxford University Press.

Partington, A. 2006. "Metaphors, motifs and similes across discourse types: Corpus-Assisted Discourse Studies (CADS) at work". In: A. Stefanowitsch and S. Th. Gries (eds.) *Corpus-based approaches to metaphor and metonymy*. Berlin and New York: Mouton de Gruyter, 267-304.

Reisigl, M. and Wodak, R. 2009. "The discourse-historical approach". In Wodak, R. and Meyer, M. (eds) *Methods of Critical Discourse Analysis* (2nd edn) London: Sage, 87-121.

Scott, M. 2011. *Wordsmith Tools.*

Woolls, D. 2012. *CFL Lexical Feature Marker.* CFL Software Ltd.

Julian Barnes' *The Sense of an Ending* and its Italian translation: a corpus stylistics comparison

Jane H. Johnson
Unversity of Bologna
janehelen.johnson@unibo.it

The distinctive style of an author is often the result of 'foregrounding' (Mukařovsky 1958; Leech 1965, 1985; Leech and Short 1981; Van Peer 1986), whereby certain linguistic elements in a work or body of works differ consistently and systematically (Mukařovsky 1958: 44) from norms represented by a particular benchmark. Leech (1985) divided foregrounding into primary, secondary and tertiary deviation according to the norm used as a benchmark. Deviation from norms may be measured by means of corpus stylistics, which provides "quantitative data in a systematic and objective way for a given phenomenon under investigation" (Mahlberg 2013: 8) and thus making it possible to "corroborate, modify and complement findings of literary critics" (ibid. 2013: 22), while also possibly revealing features of long texts which might otherwise remain invisible (Stubbs 2005: 22).

Previous corpus stylistics studies have focussed on foregrounding in the shape of highly frequent words and word clusters in order to identify particular authorial style (e.g. Hoover 2002; Johnson 2009). Other studies have made use of corpus stylistics methods to extract typical phraseological units (e.g. Starcke 2006; Fischer-Starcke 2009, 2010), or clusters, which were then categorised into local textual functions (e.g. Malhberg 2007a; 2013), as well as to identify and analyse typical collocates (e.g. Hori 2004). Corpus stylistics methods have also been useful in quantifying typical patterns of speech and thought presentation in order to judge the degree and effect of narratorial involvement in a novel (Semino 2004; Semino and Short 2004). Similarly, frequency profiles have been extracted to investigate the idiolects of particular characters (e.g. Burrows 1987; Culpeper 2002), while semantic prosodies have also been explored (e.g. Adolphs and Carter 2002; O'Halloran 2007).

The above-mentioned studies are examples of corpus stylistic analyses of novels in their original language. However explorations of the literary works in translation may similarly be performed profitably using corpus stylistics methods, often in order to highlight similarities and differences between the Source and Target Text. Indeed, if we hold that translators need to recreate predominant stylistic features of the Source Text in order to maintain stylistic or translational equivalence (Popovic 1976), it follows that the functional equivalent (ibid. 1976: 6) of any significant foregrounding in the original text should be recreated in the Target Text.

Recent corpus stylistic studies which have taken into account similarities and differences between Source and Target Text include Mahlberg (2007b), who looked at textual functions of lexis in a Dickens novel and in its German translation, Čermáková and Fárová (2010), who compared keywords in Harry Potter novels in English, Czech and Finnish and Bosseaux (2007), who focussed on point of view in various French translations of novels by Virginia Woolf. Other stylistic studies have compared novels in Italian by Grazia Deledda with their English translations, focussing on the characteristic figurative trope of simile (Johnson 2014), the use of certain deictic references in creating a specific point of view (Johnson 2011), and the effects of different renderings of salient Mental processes in the Source and Target Texts (Johnson 2010).

The focus of the present study is the tracing of elements of foregrounding in a similar contrastive framework involving both Source and Target texts. More specifically, the study uses corpus stylistics methods to explore to what extent the same elements of style identified in the original English of the novel are evident in its translation into Italian.

A recent corpus stylistic analysis (Shepherd and Berber Sardinha 2013) of Julian Barnes' Booker Prize-winning novel 'The Sense of an Ending', highlighted the keyness of various linguistic structures indicating uncertain impressionistic perceptions ('*seemed*', '*as if*' clauses, repetition of '*perhaps*'), noting that the second part of the novel deals with abstract concerns such as '*life*', and drawing attention also to a number of repeated n-grams which played a significant part in the structure of the novel.

The present study takes as its cue such earlier studies of Barnes' style to further examine aspects such as point of view (Simpson 1993) and the significance of memory in the novel as emerging from its Italian translation. The study compares the two parts of the novel with each other in both the original English and the Italian translation, thus applying notions of tertiary or internal deviation and using corpus stylistics methods to identify 'good bets' (Leech 2008: 164) to follow up for subsequent qualitative analysis. For example, it was found that key clusters belonging to different semantic groups emerged when the two parts of the novel were compared in the Italian version, in relation to the results of the same procedure in the Source Text. Those belonging to the category of 'uncertain impressionistic perceptions' (Shepherd and Berber

Sardinha 2013), frequent in the Source Text, were minimally present in the key clusters of the Target Text. Instead, a number of Mental processes figured among the keywords of Part Two when compared with Part One of the Target Text, whereas this was not the case in the Source Text. The study considers these and other findings, also in the light of broader issues such as:

- how far should the translation resemble the original anyway;
- to what extent is it the limitations of corpus stylistics that influence what we find;
- to what extent the particular target language shapes the findings of such a study.

References

Adolphs, S. and Carter, R. 2002. "Point of view and semantic prosodies in Virginia Woolf's To the Lighthouse". *Poetica*, 58: 7-20.

Barnes, J. 2011. *The Sense of an Ending*. London: Vintage.

Barnes, J. 2013. *Il Senso di una Fine* [2011 *The Sense of an Ending*]. Tr. S. Basso. Torino: Einaudi.

Bosseaux, C. 2007. *How does it feel? Point of view in translation*. Amsterdam/NY: Rodopi.

Burrows, J. F. 1987. *Computation into criticism. A study of Jane Austen's novels and an experiment in method*. Oxford: Clarendon.

Čermáková, A. and Fárová, L. 2010. "Keywords in Harry Potter and their Czech and Finnish translation equivalents". In F. Čermák, A. Klégr, P. Corness (eds.) *InterCorp: Exploring a Multilingual Corpus*. Studie z korpusové lingvistiky, svazek 13. Praha: NLN.

Culpeper, J. 2002. "Computers, language and characterisation: An analysis of six characters in Romeo and Juliet". In U. Melander-Marttala, Östman, C., Kytö M. (eds.), *Conversation in Life and in Literature*. Uppsala: Universitetstryckeriet.

Fischer-Starcke, B. 2009. "Keywords and frequent phrases of Jane Austen's Pride and Prejudice: a corpus stylistic analysis", *International Journal of Corpus Linguistics* 14: 492-523.

Fischer-Starcke, B. 2010. *Corpus linguistics in literary analysis: Jane Austen and her contemporaries*. London: Continuum.

Hoover, D.L. 2002. "Frequent word sequences and statistical stylistics", *Literary and Linguistic Computing*, 17(2): 157-80.

Hori, M. 2004. *Investigating Dickens' Style. A Collocational Analysis.* Basingstoke: Palgrave MacMillan.

Johnson, J.H. 2009. "Towards an identification of the authorial style of Grazia Deledda: a corpus-assisted study", *CeSLiC Occasional Papers*, ISSN: 2038-7954 http://amsacta.cib.unibo.it/00002678/

Johnson, J.H. 2010. "A corpus-assisted study of parere/sembrare in Grazia Deledda's Canne al Vento and La Madre. Constructing point of view in the Source Texts and their English translations", in J. Douthwaite and K. Wales (eds.), *Stylistics and Co. (unlimited) – the range, methods and applications of stylistics*, Textus XXIII: 283-302.

Johnson, J.H. 2011. "The use of deictic reference in identifying point of view in Grazia Deledda's Canne al Vento and its translation into English", in *Target* 23(1): 62-76.

Johnson, J.H. 2014. "'...like reeds in the wind'. Exploring simile in the English translations of Grazia Deledda using corpus stylistics". In D.R. Miller and E. Monti (eds.) *Tradurre Figure/ Translating Figurative Language*. Bologna: Bononia University Press.

Leech, G.N. 1965. "'This bread I break': Language and Interpretation", Review of English Literature, 6.2 London: Longmans, Green.

Leech, G.N. 1985. "Stylistics". In T. van Dijk (ed.) *Discourse and literature: new approaches to the analysis of literary genres*. Amsterdam/Philadelphia: John Benjamins.

Leech, G.N. 2008. *Language in Literature: Style and Foregrounding*. Harlow: Pearson Longman.

Leech, G. N. and Short M. H. 1981. *Style in fiction: a linguistic introduction to English fictional prose*. London: Longman.

Mahlberg, M. 2007a. "Corpus stylistics: bridging the gap between linguistic and literary studies". In M. Hoey, M. Mahlberg, M. Stubbs, W. Teubert (eds.) *Text, discourse and corpora*. London: Continuum.

Mahlberg, M. 2007b. "Corpora and translation studies: textual functions of lexis in Bleak House and in a translation of the novel into German". In M. Gatto and G. Todisco (eds.) *Translation. The State of the Art/ La Traduzione. Lo stato dell'arte.* Ravenna: Longo.

Mahlberg, M. 2013. *Corpus stylistics and Dickens's fiction*. New York/London: Routledge.

Mukařovsky, J. 1958. "Standard language and poetic language". In P. L. Garvin (ed. and tr.) A *Prague School Reader on Esthetics, Literary Structure and Style*. Washington D.C.: American University Language Centre.

O'Halloran, K. 2007. "Corpus assisted literary evaluation", *Corpora* (2): 33-63.

Popovic, A. 1976. *A Dictionary for the Analysis of Literary Translation*. Edmonton: Department of Comparative Literature, University of Alberta.

Semino, E. 2004. "Representing Characters' Speech and Thought in Narrative Fiction: A Study of England England by Julian Barnes". *Style*, 38, 4: 428-451.

Semino, E. and Short, M.H. 2004. *Corpus Stylistics. Speech, writing and thought presentation in a corpus*

of English writing. London: Routledge.

Shepherd, T.M.G. and Berber Sardinha, T. 2013. A Rough Guide to doing Corpus Stylistics. Matraga, rio de janeiro, v.20, n.32, Jan/Jun. 2013.

Simpson, P. 1993. *Language, Ideology and Point of View*. London/New York: Routledge.

Starcke, B. 2006. "The phraseology of Jane Austen's Persuasion: phraseological units as carriers of meaning". *ICAME Journal*, 30: 87-104.

Stubbs, M. 2005. "Conrad in the computer: examples of quantitative stylistics methods", *Language and Literature*, 14 (1): 5-24.

van Peer, W. 1986. *Stylistics and Psychology: Investigations in Foregrounding*. London: Croom Helm.

Nineteenth-century British discursive representations of European countries: Russia and France in *The Era*

Amelia Joulain-Jay
Lancaster University
a.t.joulain@lancaster.ac.uk

1 Introduction

Nineteenth-century Europe is considered to have been exceptionally peaceful. This peace was maintained by a complex system of interrelationships between nations which emerged as a consequence of the Congress of Vienna. The fragile balance of power meant that although skirmishes and direct military confrontations still occurred among Europe's Great Powers, they did not side with one another in fixed configurations – Britain could side with Russia on one occasion and be against her the next (Schroeder 2000:159-160,164).

The recent digitization of a very large number of Victorian periodicals offers linguists and historians an invaluable new perspective on the way that British public discourse narrated and represented its allies and enemies. Here, I focus on one publication, *The Era*, and on Britain's two primary rivals, Russia and France. I explore frequent phraseologies involving *France* and *Russia*, searching for common and diverging strategies of representation over a sixty year period, 1840-1899.

2 The corpus: *The Era* (1838-1900)

The Era is one of 48 newspapers from the nineteenth century recently digitized by the British Library (see King 2005). A Sunday paper, known as the 'leading theatrical journal of the Victorian period' (Brake & Demoor 2009:206), it was primarily intended for an audience of publicans; it contains a range of types of articles from domestic and foreign news to sports, advertisements, and articles about entertainment (mostly theatre and music-hall).

The size and generic make-up of the paper shift over the sixty-year period. Issues from the 1840s contain less than 10,000 words per issue; issues from the 1890s contain up to 180,000 words. Early issues contain around 37% news, 16% advertisements, and 15% articles about entertainment; in later issues, the proportions are respectively around 13%, 45% and 41%.

3 Distributions of *Russia* and *France*

Russia is mentioned overall less often than *France*, with the distribution of raw and relative frequencies

shown in Figure 1. Figure 1 shows that for both countries patterns of peaks and troughs are evident; some of them aligned (e.g. 1869-71), most of them not. It is tempting to look at these peaks and hypothesise that they are driven by wars involving these countries at the times in question. This initial interpretation is congruent with at least some parts of the graphs (Russia's peak in 1854-56 coincides with the Crimean War; France's 1870 peak coincides with the Franco-Prussian war).

However, examining patterns of collocation between these countries' names and *war* or words semantically related to war (G3, in the system of semantic tags used by the USAS semantic analysis system, Rayson et al. 2004) reveals that in no year do *France/Russia* co-occur with *war* more than 10% of the time, nor with the G3 category more than 20% of the time. This is even the case if the collocation span is widened to 20 tokens around the node. Figures 2 and 3 hence suggest that the factors which drive changes in frequencies of *France/Russia* vary over time. The simultaneous raw-frequency peaks for *France* and for its co-occurrences with *war* and G3 in 1870 are best explained as driven by the Franco-Prussian War; likewise Russia's 1854-56 peak reflects the Crimean War. But France's major 1892 peak and Russia's minor peaks in 1844/1883 cannot be attributed to discussions of actually-occurring wars. These observations suggest that these countries are referenced in multiple contexts, in addition to discussion of wars involving these countries.

Although, as we see from Fig. 1, the raw number of mentions for both countries oscillates around the same point across the period 1838-1900 (around 400 mentions of *France* per year, and around 100 mentions of *Russia* per year), the relative frequency of both country names decreases over that time-period, most noticeably for *France*. This may be explained by the change in generic make-up (cf. section 2): the increase in size of issues was mostly achieved by including more advertisements and articles about entertainment, which contain relatively few mentions of *Russia* and *France* relative to the (non-enlarged) part of the issues that consisted of news.

4 Analysing phraseologies for *Russia* and *France*

To analyze the phraseologies associated with *Russia* and *France* in *The Era*, I adopted Sinclair's iterative methodology (1991:84), examining 120 concordance lines (10 per country per decade) per iteration. I focused on grammatical, and secondarily semantic, relationships involving *Russia* and *France*, grouping similar phraseologies into categories. The set of categories devised from the first sample was then applied to further cycles of samples until a sample was reached in which no further modifications/extensions to the analysis were necessary.

The resulting framework (table 1) consists of three overarching categories, each encompassing four or five specific phraseologies. The phraseologies are represented using slot-pattern schemas; slot labels are capitalized if they could be instantiated by a range of tokens (e.g. **COUNTRY** could be instantiated as *Russia*, *France*, etc.) or lowercased if they represent a slot filled by a specific form. Optional slots are bracketed.

Five phraseologies (1.1-1.5) were categorized as 'personifying' because they represent the countries as fulfilling some semantic role (prototypically *agent*) associated with (conscious) human beings.

Five phraseologies (2.1-2.5) were categorized as 'locational' because they represented the countries in their geographical sense, as a *location*, *source*, or *goal*. Four rare phraseologies (3.1-3.4) were labelled 'specialized' due to being associated with a narrow set of contexts; for example, '**COUNTRY NUMBER'** only occurs in lists of places and numbers such as sporting results or reports of goods shipped.

Three types of instances were excluded from the analysis: cases from passages where poor-quality OCR precludes identification of the phraseology; cases where the country names are part of titles of articles or shows; and cases where *France/Russia* does not refer to the country (but is instead the name of a person, horse, etc.).

5 Representations of *Russia* and *France*

The representations of Russia and France have more in common than they have differences. Most strikingly, the phraseologies in use shift over time as the publication's genre-make-up evolves, so that the two last decades are markedly different from the others. Whereas mentions of the countries within the 'personifying' and 'locational' patterns account for almost all the concordance lines in the early decades, in the last two decades, the number of **(TROUP/VENUE) (TOWN) COUNTRY** phraseologies ('specialized' because found almost exclusively in addresses within advertisements) rises sharply. From no occurrences in the samples for the first four decades of *Russia*, and just one in those of *France*, it rises to 25% for *Russia* and 45% for *France* in the last two decades.

No	Schema	**Example** (OCR errors not corrected)
1.1	NOUN_PHRASE ATTRIBUTIVE_ PREPOSITION COUNTRY	"The complaints of the enormous intrigues of Russia are becoming universal." (23/10/1842)
1.2	COUNTRY=VERB_COMPLEMENT	"if any European power opposes Russia in her projects" (10/6/1866)
1.3	COUNTRY PREDICATE	"Russia is greedy for Batoumn" (07/7/1878)
1.4	COUNTRY to VERB	"Hebelievedtherewas no intention on the part of France to interfere" (20/6/1847)
1.5	NOUN_PHRASE between COUNTRY and COUNTRY	"The Excitement in Constantinople , caused by the late War between Russia and Turkey…" (08/12/1878)
2.1	NOUN_PHRASE LOCATIONAL_ PREPOSITION COUNTRY	"…deals with life in Russia during the Crimean War" (15/8/1896)
2.2	VERB (LOCATIONAL_PREPOSITION) COUNTRY	"We are the only Agents who have travelled over India , (…) Germany , Russia" (12/6/1886)
2.3	In COUNTRY	"We expect a very considerable rise in French shares, especially in the Northern of France, which line will divide 8 per cent." (06/2/1848)
2.4	ADJECTIVE LOCATIONAL_ PREPOSITION COUNTRY	"Bobi is said to be a very favourite comnedy in Russia" (19/9/1896)
2.5	INSTITUTION of COUNTRY	"the Emperor of RUSSIA , most respectfully solicits from the public an Inspection of his extensive STOCK of WATCHES" (13/5/1849)
3.1	(TROUP/VENUE) (TOWN) COUNTRY	"Engagement with CIRCUS CINISELLI , ST . PETERSBURG , RUSSIA , on their Three Horizontal Bars" (17/2/1883)
3.2	COUNTRY DATE	"France , Mlay 25th , 1888 . My dear Goddard , Your letter of 2ed has reassured me…" (23/6/1888)
3.3	COUNTRY NUMBER	"Tire total quantity amounted to 2,689,000 bottles , which were thus distri-buted : England and British India , 467,000 ; Russia and Poland , 502,000 ;" (24/10/1852)
3.4	COUNTRY NOUN	"Penetrating Hair Brushes, with the durable unbleached Russia Bristle, which do not soften like common hair." (05/5/1850)

Table 2. Phraseologies occurring with *Russia* and *France* in *The Era* (1840-1899)

Distinctions between the representations of the countries are, however, noticeable in the details of the specific 'personifying' and 'locational' phraseologies associated with each country. When *Russia* occurs in an 'personifying' phraseology, it is overwhelmingly (in 3/4 of cases) in a **NOUN_ PHRASE ATTRIBUTIVE_PREPOSITION COUNTRY** configuration, e.g. 'the complaints of the enormous intrigues of Russia are becoming universal' (*The Era*, 23/10/1842). In contrast, *France* appears within a more diverse range of phraseologies. The most common, occurring in about a third of cases, is **COUNTRY PREDICATE**, e.g. 'France was going to war' (*The Era*, 09/1/1859). This result suggests a subtle difference in the amount of agency – or, to put it another way, ability to exert power – assigned to the two countries.

In terms of 'locational' phraseologies, in over half the cases, *Russia* occurs in an **INSTITUTION of COUNTRY** pattern, e.g. 'the emperor of Russia most respectfully solicits from the public an Inspection of his extensive stock of watches' (*The Era*, 13/5/1849), whereas *France* tends to occur either in **NOUN_PHRASE LOCATIONAL_ PREPOSITION COUNTRY** patterns, e.g. '[the ship] brings (…) 297 passengers for England and France' (*The Era*, 06/6/1858), or in **VERB (LOCATIONAL_PREPOSITION) COUNTRY** patterns, e.g. 'a pretty American girl who had been educated in France' (*The Era*, 18/11/1877). This suggests (unsurprisingly) that France is represented as a place from and to which things and people may move more than Russia, which is predominantly constructed as a setting in which events being described are played out.

Acknowledgement

This research is part of the ERC-funded *Spatial Humanities: Texts, GIS and Places* project at Lancaster University[61].

[61] http://www.lancaster.ac.uk/fass/projects/spatialhum.wordpress/

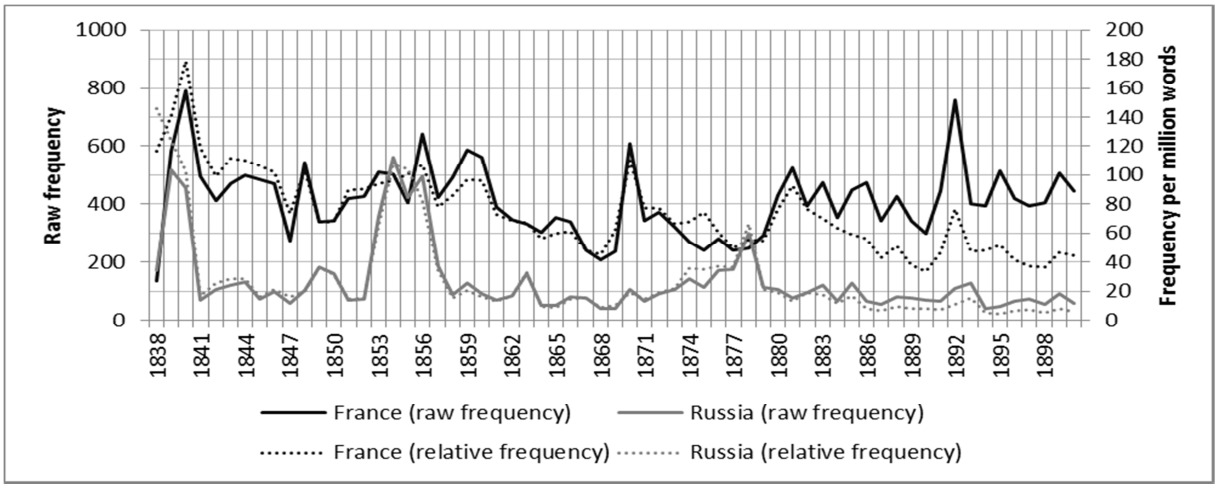


Figure 1. Raw and relative frequencies per year of *Russia* and *France* in *The Era* (1838-1900).

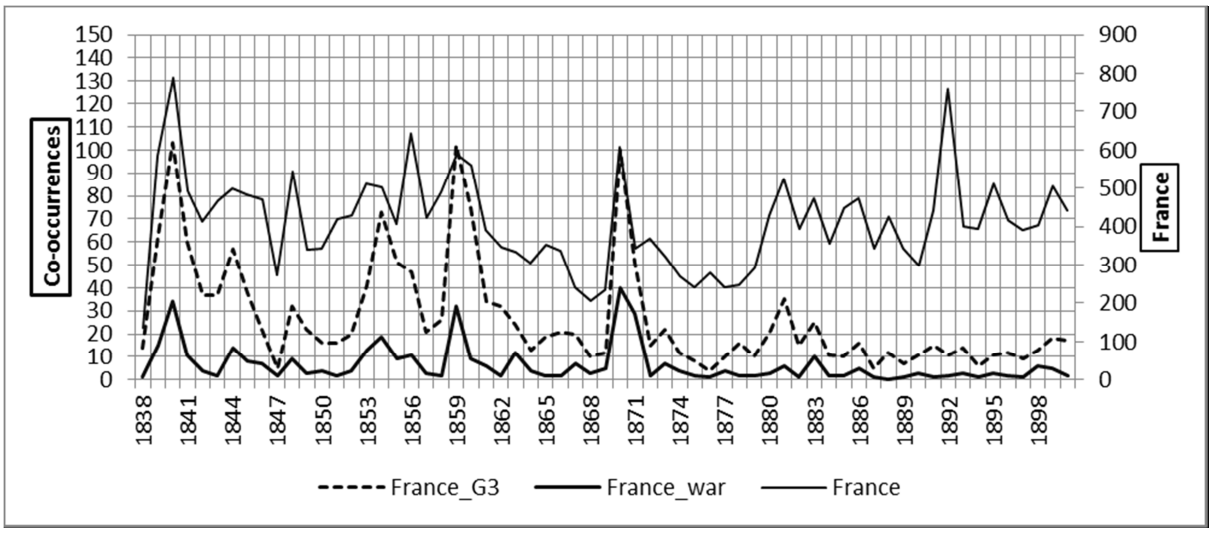


Figure 2. Raw frequency per year for *France*, *France* co-occurring with *war* within 20 words, and *France* co-occurring with words tagged G3 within 20 words in The Era (1838-1900).

Figure 3. Raw frequency per year for *Russia*, *Russia* co-occurring with *war* within 20 words, and *Russia* co-occurring with words tagged G3 within 20 words in The Era (1838-1900).

References

Brake, L. and Demoor, M. (eds.) 2009. *Dictionary of Nineteenth-Century Journalism in Great Britain and Ireland*. London: Academic Press and the British Library.

King, E. 2005. "Digitisation of Newspapers at the British Library". *The Serials Librarian* 49 (1-2): 165-181.

Rayson, P., Archer, D., Piao, S. L. and McEnery, T. (2004). "The UCREL semantic analysis system". In *Proceedings of the workshop on Beyond Named Entity Recognition for NLP tasks* in association with *4th International Conference on Language Resources and Evaluation* (LREC 2004), 25th May 2004, Lisbon, Portugal, 7-12. Paris: European Language Resources Association.

Schroeder, P. W. (2000). "International politics, peace, and war, 1815-1914". In Blanning T. C. W. (ed.) *The Nineteenth Century: Europe 1789-1914*. Oxford: Oxford University Press.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

"All our items are pre-owned and may have musty odor": A corpus linguistic analysis of item descriptions on eBay

Andrew Kehoe
Birmingham City University
andrew.kehoe@bcu.ac.uk

Matt Gee
Birmingham City University
matt.gee@bcu.ac.uk

1 Introduction

This paper presents the first large-scale corpus linguistic analysis of the popular online auction site eBay. Founded in 1995, eBay provides registered users with a marketplace for the sale of a wide range of goods. The site has 152 million active users worldwide, with 800 million items listed for sale at any given time (all statistics from eBay Inc. 2014).

Although it is often thought of as an auction site where members of the public can sell unwanted gifts and household clutter to the highest bidder, eBay contains an increasing proportion of fixed price (non-auction) items, and 80% of all items are new rather than used. These include goods offered for sale by small businesses using eBay as their main 'shop window', as well as large retailers using the site as additional online sales channel.

Crucially, all sellers on eBay, whether they be individuals or large companies, are encouraged to describe in their own words the items they list for sale. In this paper we analyse a large sample of these item descriptions, exploring linguistic norms on eBay as well as linguistic variation between sellers and types of item.

2 Corpus composition

In the first part of the paper we describe our corpus compilation process. Items on eBay are offered for sale in categories, of which there are 35 at the top level ('Computers/Tablets & Networking', 'Sporting Goods', 'Baby', etc.), each with its own sub-categories. For this study we have built a corpus of item descriptions across all categories on eBay's UK website,[62] one of 25 country-specific sites.

Compiled over four months using our bespoke WebCorpLSE crawling tools (Kehoe & Gee 2009), our corpus contains over 400,000 item descriptions totalling 100 million words. We included only completed items (items that had closed at the time of our crawl) and, in addition to the textual descriptions, we recorded item category and sale price. All textual descriptions have been part-of-speech tagged using TreeTagger (Schmid 1994).

[62] http://www.ebay.co.uk/

3 General eBay lexicon

We begin our linguistic analysis by presenting our initial attempts to compile a general eBay lexicon through the examination of word frequencies across item categories. We have found that there are certain core words relating to eBay processes and protocols which appear consistently across all categories: *item, buyer, seller, payment, paypal, shipping, feedback,* etc.

Furthermore, we have found that a high concentration of these words tends to be indicative of boilerplate: standard text that appears on all listings by a particular seller. This text is found in particular on eBay listings by companies rather than individuals, with slight variations between sellers. By identifying boilerplate at an early stage, we are able to focus on more interesting linguistic examples in the remainder of the paper.

4 Linguistic variation

Our primary focus is on linguistic differences between the descriptions of items in the various eBay categories, with particular reference to the adjectives used by sellers to describe items for sale in these categories. For this, we adopt a keywords approach (Scott 1997) to compare sub-corpora (categories) with one another.

Although the core eBay-related words appear consistently across categories and while there are obvious differences in the frequent topic-related words in each category (primarily nouns), we find significant differences in adjective use between categories too.

Our first case study concerns the words used to describe used items, which vary from *used* itself to *second-hand*, *pre-owned, pre-loved*, etc. We also examine adjective use by individual sellers. We have found that the boilerplate text appearing on listings from some sellers contains detailed explanations of how that seller defines particular terms (e.g. "A product is listed as 'Used' or 'Unsealed' when the manufacturers seal on the product box has been broken"). Different terms are used differently by different sellers, and we explore this through collocational analysis.

Our second case study concerns 'fake' items. We have found only a limited number of circumstances in which sellers describe their own items using the word *fake*: *fake fur, fake eyelashes*, etc. A much more common scenario is for sellers to warn buyers about fake items being sold by rival sellers ("Quit worrying about buying fake autographs on ebay and buy a REAL one here!"). We also find that many eBay categories have their own particular euphemisms for describing fake items, e.g. *non-original, generic,* and *compatible* in Computing. We investigate this phenomenon in depth, drawing examples from the corpus and carrying out collocational analyses.

5 Adjective use by price band

A further dimension in our analysis is price. We have produced a price distribution for all items in our corpus and carried out a keyword comparison between the cheapest 25% of items and the most expensive 25%. We present findings from this analysis, including a discussion of the adjectives associated more frequently with items in the cheap (e.g. *lightweight, plastic, acrylic, ex-library*) and expensive (e.g. *heavy, steel, leather, pristine*) categories.

6 Summary

Throughout the paper we explain that a deeper understanding of the language of online selling is vital as e-commerce continues to grow worldwide. Although there are commercial companies such as Terapeak[63] offering analyses of general trends on eBay, we are not aware of any in-depth academic analyses of the language of eBay or other e-commerce sites. In our paper we give examples of how corpus linguistic techniques can be applied to the study of this increasingly important social phenomenon, and suggest how our techniques could be used to improve the indexing and search functions on sites like eBay.

References

eBay Inc. 2014. *eBay Marketplace Fast Facts At-A-Glance (Q3 2014) – Shareholders' Report*: http://bit.ly/eBayInc2014

Kehoe, A. & M. Gee. 2007. "New corpora from the web: making web text more 'text-like'". In P. Pahta, I. Taavitsainen, T. Nevalainen and J. Tyrkkö (eds.) *Towards Multimedia in Corpus Studies,* electronic publication, University of Helsinki: http://www.helsinki.fi/varieng/journal/volumes/02/kehoe_gee/

Schmid, H. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees". *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Scott, M. 1997. "PC Analysis of Key Words – and Key Key Words". *System* 25 (1), 1-13.

[63] http://www.terapeak.com/

Corpus-based analysis of BE + *being* + adjectives in English

Baramee Kheovichai
Silpakorn University
kiao_ra@yahoo.com

1 Introduction

While grammar books often write that stative verbs such as 'BE' cannot be used in progressive aspects, one can find examples of verb 'be' used in progressive aspects as in "You are being stupid again" (Kennedy, 2003: 232). According to Kennedy (2003), the adjectives used in this context are dynamic adjectives where the grammatical subject is in control of the state and thus these adjectives license this pattern. This explanation is broad and does not account for the meaning of the grammar pattern.

As pointed out by, to name a few, Hunston and Francis (2001) and Hoey (2007), there is a close tie between grammar and lexis,. As Hunston and Francis state, a grammatical pattern is associated with a meaning or many meanings. They observe that lexis generally occurs in the pattern that has related meaning to it, thereby displaying a co-selection between lexis and grammar. As a result, it is of particular interest to this study to see what meanings are associated with this pattern and what adjectives can be part of this pattern. Furthermore, other linguistic features which can also influence the meanings and uses of this grammar pattern are investigated.

2 Purposes

This paper aims to investigate the pattern BE + being + Adjectives in the British National Corpus (http://bncweb.lancs.ac.uk). The research questions are: 1) what adjectives can go in the slot, 2) what are the meaning groups of these adjectives, 3) what tense is more frequently used in this pattern, 4) what are the grammatical subjects of this pattern and 5) what categories of engagement occur in this pattern. This work hopes to shed light on the meaning of this grammar pattern resulting from the co-selection of grammar and lexical features associated with this pattern based on a corpus-based analysis of authentic language. The analysis was based on the British National Corpus which contains a large size of data and allows for part-of-speech search, thereby facilitating the retrieval of this pattern in the corpus.

This paper considers the adjectives that can occur in this pattern, the tense, the types of grammatical subject and the linguistic features associated with engagement (Martin & White, 2005). The adjectives can indicate the qualities or characteristics associated with this pattern, the analysis of grammatical subjects can indicate the objects of referred to in this pattern; that is, what kind of people or things are often talked about, using this pattern. The tense can indicate if this pattern is used for current, past or future events. Furthermore, the framework of engagement can shed light on the interactive aspect of this pattern. These variables can contribute to the meanings associated with the pattern.

3 Methodology

The research procedures are as follows. First, a search term _VB* (n't)? (not)? being (_AV*)? (_AV*)? (_AV*)? _AJ* was entered into the BNC webpage. The resulting outputs were exported into Microsoft Excel. They were then manually coded. The adjectives were sorted into categories according to the Appraisal Framework (Martin & White, 2005). The classification of the subject was adapted from Gries (2006). The framework of engagement (Martin & White, 2005) was used to classify the engagement features. This study only focuses on instances where BE + being + adjectives is used in progressive aspect and thus when this pattern is used in a pseudo-cleft sentence, the instances were excluded from the analysis.

4 Results and discussion

There are 1,239 instances of this pattern in total. The categories of evaluative meanings, example adjectives and their frequency is shown in Table 1. In terms of the broad categories, the analysis indicates that Judgement has the highest frequency (993 instances), followed by Affect (128 instances) and Appreciation (118 instances). Therefore, it is apparent that this phraseological pattern is more strongly associated with Judgement than others. Regarding the subcategories, for Judgements the three most frequent categories are: 1) –propriety (215 instances), 2) –capacity (194 instances) and 3) +propriety (158 instances). These three are in fact the most frequent subcategories and thus indicate that this phraseological pattern is related to propriety and ability. The three most frequent subcategories of Affect are: 1) security (31 instances), 2) dissatisfaction (29 instances) and 3) insecurity (24 instances). The three most frequent categories of Appreciation are: 1) +composition (27 instances), 2) –composition (27 instances) and 3) –reaction (22 instances). Overall, it seems that the pattern is more associated with negative meaning or has negative semantic prosody.

	Evaluative meanings	Adjectives	No.
Judgement	+Normality	Normal	1
	-Normality	Unusual	2
	+Capacity	Clever, sensible, successful	67
	-Capacity	Silly, unreasonable, ridiculous	194
	+Tenacity	Cautious, careful, brave	110
	-Tenacity	Irresponsible, hasty, aggressive	147
	+Veracity	Honest, frank, truthful	53
	-Veracity	Tactless, coy, disingenuous	46
	+Propriety	Nice, kind, generous	158
	-Propriety	Unfair, rude, horrible	215
Affect	Security	Reassuring, positive, assertive	31
	Insecurity	Paranoid, cagey, pessimistic	24
	Satisfaction	Appeasing, complacent, attentive	7
	Dissatisfaction	Dismissive, hysterical, grumpy	29
	Happiness	Happy, cheerful, euphoric	6
	Unhappiness	Dangerous, gloomy	2
	Inclination	Affectionate, intrigued, admiring	13
	Disinclination	Choosy, frightened, indifferent	16
Appreciation	+Composition	Direct, exact, consistent	27
	-Composition	Evasive, cryptic, inconsistent	27
	+Reaction	Amusing, funny, pleasant	13
	-Reaction	Disgusting, noisy, unpleasant	22
	+Valuation	Effective, wonderful, remarkable	13
	-Valuation	Futile, under-utilised	2
	Others	Italian, cloudy, windy	14
	Grand Total		1239

Table 1

The majority of this pattern is in the present simple tense (747 instances), followed by the past simple tense (491 instances). There is no instance where this pattern is used in the future tense. As such, this indicates that this pattern is used in reference to current events and in some cases past events.

In terms of the grammatical subjects, a third person human is most frequently used as a subject (634 instances), followed by first person (313 instances) and second person pronoun (184 instances). As a consequence, this pattern is most often used to talk about the behaviors or characteristics of other people or the speaker.

In terms of engagement, the phraseological pattern is more frequently oriented to heterogloss (715 instances) than monogloss (524 instances). That is, it is more strongly associated with interactive features, incorporating authorial stance than with definiteness. Within heterogloss, there are subcategories and the most frequent one is disclaim (323 instances). Consequently, this pattern is associated with contrast and negation.

This study has shown that the pattern BE + being +Adjectives is more complicated than only instances of dynamic adjectives used in progressive aspect. Given that adjectives in this pattern frequently refer to impropriety, incapacity and other negative meanings, it could be argued that this pattern has a negative semantic prosody. This pattern is used for present event and it often refers to other people. Moreover, this pattern has engagement features and is especially used to disclaim. This paper has hopefully cast further light on the meaning and function of the pattern BE + being + Adjectives through analysis of collocation based on authentic language use.

References

Gries, S. T. (2006). "Corpus-based methods and cognitive semantics: The many senses of to run". In S. T. Gries & A. Stefanowitsch (Eds.), *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis* (pp. 57–100). Berlin/New York: Mouton de Gruyter.

Hoey, M. 2007. "Lexical priming and literary creativity". In M. Hoey, M. Mahlberg, M. Stubbs, & W. Teubert (Eds.), *Text, discourse and corpora: Theory and analysis* (pp. 7–29). London and New York: Continuum.

Hunston, S., & Francis, G. 1999. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam and Philadelphia: John Benjamins Publishing.

Kennedy, G. 2003. *Structure and Meaning in English: A Guide for Teachers*. London: Pearson Education Limited.

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation*. Palgrave Macmillan: Great Britain.

DIACRAN: a framework for diachronic analysis

Adam Kilgarriff
Lexical Computing Ltd.
Adam.kilgarriff@sketchengine.co.uk

Ondřej Herman
Lexical Computing Ltd., Masaryk University
Ondrej.herman@sketchengine.co.uk

Jan Bušta
Lexical Computing Ltd., Masaryk University
Jan.busta@sketchengine.co.uk

Vojtěch Kovář
Lexical Computing Ltd., Masaryk University
Vojtech.kovar@sketchengine.co.uk

Miloš Jakubíček
Lexical Computing Ltd., Masaryk University
Milos.jakubicek@sketchengine.co.uk

1 Introduction

Many of the questions that linguists want to explore concern language change, or diachronic analysis. We present Diacran, an implemented system for corpus-based diachronic analysis.

We view diachronic analysis as a special case of keyword-finding. In keyword-finding we want to find the words (or terms, or collocations, or grammatical structures) that are most characteristic of one text type (or dataset, or corpus) in contrast to another. In diachronic analysis, we usually want to start by finding the words (or terms, etc; hereafter we say just 'word') that have changed most over time. The ingredients for keyword analysis are two corpora and a formula for ranking how interesting each word is. For Diacran, the ingredients are a corpus with at least three 'time-slices' - that is, with documents dated according to at least three different points in time so the corpus can be sliced into three or more subcorpora, each associated with a different time - and, again, a ranking formula.

As in keyword analysis, the challenge for the computational linguist is of getting the 'best' list, where 'best' means the list of (say) the top 500 items, with the largest numbers of items judged interesting (from a text-type, or diachronic, point of view) by a human expert.

The method is this. First we divide the corpus into subcorpora, one for each time slice. Then we normalize the frequency for each word in each time slice, to give frequencies per million words.[64] We then plot a 'best fit' graph (straight line), for each word, of change over time, using standard techniques such as linear regression and Theil-Sen gradient estimation.

The 'most interesting' of these graphs have three characteristics:

1 high gradient (positive or negative) of the line
- **1.1** because we are most interested in words that have changed a lot

2 high correlation between the regression line and source graph
- **2.1** because we are most interested in words that have changed and stayed changed, not bounced around
- **2.2** indicates high credibility and can be computed as statistical significance as defined for particular regression methods

3 high frequency of the word overall
- **3.1** because we are more interested in words where the frequencies in the (say, five) time slices are <100, 200, 300, 400, 500> rather than <1, 2, 3, 4, 5>. The latter is likely just to be noise, whereas for the former, we have ample evidence of a systematic and substantial change.

We then combine the scores on these three factors, to give an overall score for each word. The words with the highest combined scores are the 'most interesting' words, to be shown to a linguist for expert consideration. Diacran is implemented within the Sketch Engine (Kilgarriff et al. 2004) and the expert is supported in their task by 'click through': they can click on an item in the candidate list to see the concordances for the word. They can also see other analyses to show how usage differs between time-slices, within the Sketch Engine, which has a wide range of analysis tools.

The approach should prove useful for various kinds of diachronic analysis. We are using COCA (Davies 2009) and a corpus of blog and newspaper feeds that we have gathered over the last ten years, as test sets.

An ideal test set would give us the 'right answers' so we knew when our system was doing a good job. We are currently searching for datasets that might support us in choosing the best ranking formula, setting thresholds, and evaluation.

64 Another option is to classify a word as present or absent in a document, and to work with counts for each word per thousand documents. This is often preferable, as we do not wish to give extra weight to a word being used multiple times in a single document. Diacran offers both options.

2 Neologisms

The highest-profile kind of diachronic analysis is neologism-finding, particularly by dictionary publishers, where the year's new words are featured in the national press. We are exploring using the set of new words, as added to a dictionary by a dictionary publisher, as the 'ground truth' of the words that our system should put high on the list.

A feature of neologism-finding, particularly for brand-new words (as opposed to new meanings for existing words) is that frequencies, even in very large corpora, will tend to be very low. A sequence of frequencies, over the last five years, of <0, 0, 1, 0, 2> for a word may count as enough to suggest a candidate neologism, that came into existence three years ago. This presents a technical challenge since there are also likely to be many typographical errors and other noise items with profiles like this. It also points to the merits of working with very large corpora, since, the larger the numbers, the better the prospects for using statistics to distinguish signal from noise.

3 Background and Related Work

The traditional way to find neologisms is 'reading and marking'. Lexicographers and others are instructed to read texts which are likely to contain neologisms – newspapers, magazines, recent novels – and to mark up candidate new words, or new terms, or new meanings of existing words. This is the benchmark against which other methods will be measured. It is a high-precision, low-recall approach, since the readers will rarely be wrong in their judgments, but cannot read everything, so there are many neologisms that will be missed.

For a dictionary publisher, one reading of 'neologism' is 'words which are not in our dictionary (yet)'. Of course words may be missing from dictionaries for many reasons, of which newness is one, sand others include simple oversight, and the dictionary's policies on derivational morphology, specialist vocabulary, and rare and archaic items. On this reading, one kind of neologism-finding program identifies all the words in a corpus (over a frequency threshold) that are not in a particular dictionary. Corpora have been used in this way to mitigate against embarrassing omissions from dictionaries since large, general corpora (for example, for English, the British National Corpus[65]) became available, in the late 1980s and 1990s. Note that this process has complexities of its own, and where the language has complex morphology, identifying the word forms not covered by the lemmas in the dictionary is far from simple.

There are some 'lexical cues' that speakers often use when introducing a word for the first time: "so-called", "defined as", "known as". In writing, the language user might put the new item in single or double quotation marks. One kind of corpus strategy for identifying neologisms looks for items that are marked in these ways. An implemented system for English, which shows these methods to be strikingly useful, is presented by Paryzek (2008).

The approach is extended for Swedish by Stenetorp (2010) who starts from lists of neologisms from the Swedish Academy and Swedish Language Council, and develops a 'supervised' machine learning system which finds features of neologisms *vs.* non-neologisms, and can then classify new items as neologism-like or not. Stenetorp uses a very large corpus of documents each with a time stamp, as do we.

O'Donovan and O'Neil (2008) present the system in use at Chambers Harrap at the time for identifying neologisms to add to the dictionary. This is of particular interest as it is a system which, in contrast to the academic ones, is used in earnest by a publisher. One component of the software suite builds a large time-stamped corpus; another, the word-tracking component (based on Eiken 2006) identifies items which have recently jumped up in relative frequency; and a third, echoing the third of our criteria above, promotes higher-frequency items so they will appear higher in the lists that lexicographers are asked to monitor.

Gabrielatos et al. (2012) present an approach to diachronic analysis similar to ours, but focusing on one specific sub-issue: what are the most useful time-slices to break the data set up into. There is usually a trade-off between data sparsity, arguing for fewer, fatter time-slices, and delicacy of analysis, which may require thinner ones. We plan to integrate the lessons from their paper into the options available in Diacran.

Acknowledgments

This work has been partly supported by the Ministry of Education of the Czech Republic within the LINDAT-Clarin project LM2010013 and by the Czech-Norwegian Research Programme within the HaBiT Project 7F14047.

References

Davies, M. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159-190.

Eiken, U. C., Liseth, A. T., Witschel, H. F., Richter, M.,

65 http://www.natcorp.ox.ac.uk

and Biemann, C. 2006. Ord i dag: Mining Norwegian daily newswire. In *Advances in Natural Language Processing* (pp. 512-523). Springer Berlin Heidelberg.

Gabrielatos, C., McEnery, T., Diggle, P. J., & Baker, P. (2012). The peaks and troughs of corpus-based contextual analysis. *International journal of corpus linguistics*, *17*(2), 151-175.

Kilgarriff, A., Rychlý, P., Smrz, P., Tugwell, D. 2004. The Sketch Engine. Proc. EURALEX. pp. 105–116.

O'Donovan, R., & O'Neil, M. 2008. A systematic approach to the selection of neologisms for inclusion in a large monolingual dictionary. In *Proc 13th Euralex International Congress* (pp. 571-579).

Paryzek, P. 2008. Comparison of selected methods for the retrieval of neologisms. Investigationes Linguisticae XVI, Poznan, Poland.

Stenetorp, P. 2010. Automated extraction of swedish neologisms using a temporally annotated corpus. Masters' thesis, KTH (Royal Institute of Technology) Stockholm, Sweden.

Corpus annotation: Speech acts and the description of spoken registers

John Kirk
Technische Universität Dresden
jk@etinu.com

To celebrate the career of Geoff Leech, I offer a paper on pragmatic corpus annotation which immediately combines two of his central research interests to which he made such original and pioneering contributions. The paper presents solutions to the challenges of speech-act annotation and, from analysis of those annotations, pragmatic characterisations of spoken registers.

Neither the basic transcription protocol nor the extended markup systems for classic spoken corpora (such as the London-Lund Corpus or the spoken components of the International Corpus of English or the British National Corpus attempt to indicate speakers' pragmatic intentions within the corpus. Yet the conveyance of pragmatic meaning is at the core of the successful conveyance of almost every utterance. Pragmatic meaning comprises several components, some explicit, some implicit, and inevitably there is a great deal of variation. It is thus a natural development to include pragmatic annotation within a corpus, and Geoff Leech greatly encouraged our approach. Towards this end, the SPICE-Ireland corpus encodes the speech act status of each utterance in the corpus, using a system that is developed from the work of Searle (1976). Searle constructs a taxonomy of what he terms (p. 10) 'the basic categories of illocutionary acts', paying attention especially to the ways in which these different acts reflect 'differences in the direction of fit between words and the world' (p. 3). Searle's taxonomy of illocutionary acts focuses on five types of speech act, labelled as *representatives*, *directives*, *commissives*, *expressives*, and *declaratives*. Searle's taxonomy is designed to illustrate systemic aspects of language use, not to encode actual examples of language in use. Nevertheless, it provides a realistic basis on which to build a system of pragmatic annotation that provides for an exhaustive and explicit categorisation of the material in a corpus. To Searle's five categories SPICE-Ireland added three of its own: 'indeterminate conversationally-relevant units' (ICU), such as feedback responses or signals such as *right*, *yes,* or *ok* which provide conversational coherence but are not uttered with an intended pragmatic function or with any other commitment in the unfolding conversation or discourse, but which are crucial to the development of the ongoing discourse; 'incomplete utterances or fragments' which are pragmatically indecipherable;

greetings, leave takings, and other interactive expressions; and 'keyings', following Goffmann (1974) for utterances involving humour or irony where speakers are not being literal or felicitous, and where normal conditions of language use do not apply.

The transcription practice in the SPICE-Ireland corpus is to mark the speech act status of an utterance with a code in angle brackets before the utterance, concluding with a backslash and the appropriate code at the end. The usual scope of an utterance for the annotation of pragmatic effect corresponds to a sentence or clause. Though it is possible to understand some strings of words as including more than one pragmatic intention, the annotation works on a principle of exclusivity, whereby only one speech act is assigned to any given string of words. Cases which appeared ambiguous are annotated on the most likely interpretation within the context of the conversation as a whole; utterances which cannot plausibly be linked to a particular function are also so marked. Grammatical notions such as clause and sentence are problematical in spoken language, and decisions on annotation were made on a combination of structural, prosodic, and discoursal features. No simple algorithm exists for determining the speech act status of an utterance; annotation is made on the basis of detailed and, it must be stressed, manual analysis of language in use.

Among its 300 texts each of 2000 words and comprising 15 discourse situations, distributed equally between Northern Ireland and the Republic of Ireland, the SPICE-Ireland corpus has 54,612 speech acts. The paper will present the raw occurrences per text category, North and South; the relativized (or normalized) frequency of those occurrences per 1,000 words, again in each text category, North and South; and thirdly the percentage of each Speech Act type per text category.

Not surprisingly, Representatives are the most frequent Speech Act type. Overall, by averaging all speech act types in each of the 15 registers, and when totaled, they amount to 65% (or almost 2 out of 3). 19% (almost 1 in 5) are Directives. Whereas this might seem intuitively satisfying, what the SPICE-Ireland corpus provides researchers with are objective, systematically- and empirically-derived distributional frequencies, quite unrivalled, as there appear to be no others with which they may be compared.

As for Directives, the highest frequency of the annotation occurs in the Demonstrations category, where speakers expect others to perform or undertake various tasks either at the time of utterance or at some time in the future. In close second is the Face-to-face Conversations category, for it is in the nature of everyday conversation to make requests, seek confirmations or clarifications, and pose questions for a wide range of purposes.

There is considerable variation in the distribution of the ICU annotation, a category created for the SPICE-Ireland annotation scheme. At the top end, it is Telephone Conversations category which has the highest score, where ICUs make up for the absence of body language in the dislocated conversations. The Face-to-face Conversations category ranks only third for use of the ICU behind the Business Transactions category, where the urging and persuading necessary to achieve agreements or undertakings may often be accompanied by ICU markers. Examples are:

(3a) <NI-TEC-P1A-098$A> <#> <rep> I 'm not even sure 2exActly when I 'll 2nEEd somebody from% </rep>
<$B> <#> **<icu>** 2Right% **</icu>**

(3b) <ROI-SCS-P2B-050$A> <#> <rep> 1MY budget target for the E-B-2R% would not then be 1incrEAsed by making further 1pAYments% </rep> <#> <rep> 1And the assets I will 1consIder 1dispOsing of% are not in the commercial semi-state 1bOdies% <,,> </rep>
<P2B-050$C> <#> **<icu>** Watch this space **</icu>**
<P2B-050$D> <#> **<icu>** Read my lips **</icu>**

With relatively low frequencies for the other speech act types, it is only the three categories of **<rep>, <dir>, and <icu>** which, in their various frequency constellations, are shown to characterise different text categories.

Moreover, the paper shows that it follows from the high frequency and percentage distribution of Representatives that, unless its occurrence is extremely high (as with the Broadcast News, Broadcast Talks, and Spontaneous Commentaries categories), it is the relative values of Directives and ICUs by which registers are discriminated.

Using these figures, it is possible to offer fresh profiles of each text category, of which the paper will briefly give indicative examples.

Like the other more conversational text type categories, Face-to-face Conversations are largely characterised by Representatives, Directives, and ICUs. The combined percentage distribution of these three speech act types accounts for 91%.

The special nature of Demonstrations leads to the frequent use in this text type of Directives, which are more common than the Representatives which dominate every other text type category. The combined percentage distribution of these two speech act types accounts for 94%.

Spontaneous commentaries, which largely focus on the provision of information and rarely allow for conversational interaction, show a very high frequency of Representatives, accounting for 93% of speech acts within the category.

In line with other conversational text type categories, Telephone conversations are largely characterised by the presence of three speech act types: Representatives, Directives and ICUs. The combined percentage distribution of these three speech act types accounts for 89.5%.

These brief profiles show that Representatives predominate in each text category except for Demonstrations, and constitute 90% or more of speech act annotations in the Broadcast News, Broadcast Talks, and Spontaneous Commentaries categories. Directives and ICUs may also be prominent, to greater or lesser degrees.

The paper refrains from speculating about these distributions on the basis of any stereo-typical text category characteristics such as spontaneity, preparedness, or scriptedness ('written to be read'), or on whether the speech is a monologue, a genuine dialogue (literally between two people), or a polylogue (between many speakers). Further research will be needed to extrapolate from these ratios and percentages any kind of correlation or cline between speech act type and text category correlation.

Speech act annotations open up many possibilities for the analysis of language in use. The validity of any analysis derived from the SPICE-Ireland corpus rests not only on the authenticity of the data but on standardisation measures such as the selection of text types, speakers, and text size, as set down by ICE protocols (cf. Greenbaum 1996). What emerges from the data, however, is both consistency and variation across text categories and the speech act types.

Although some findings reported in the paper may lend themselves to comparison with text category characteristics made on qualitative or impressionistic grounds in the past, no other studies comprise such a broad range of spoken text categories (as an ICE corpus so conveniently facilitates) which have received a pragmatic profiling along the present qualitative lines, or with which the present quantitative results may be compared.

The Asian Corpus of English (ACE): Suggestions for ELT policy and pedagogy

Andy Kirkpatrick
Griffith University
Brisbane
a.kirkpatrick@griffith.edu.au

Wang Lixun
Hong Kong Institute of Education
lixun@ied.edu.hk

ACE is a corpus of some one million words of naturally occurring data of English used as a spoken lingua franca by Asian multilinguals. Nine data collection teams across Asia worked together to collect and transcribe the corpus. ACE was inspired by VOICE, the Vienna Oxford International Corpus of English, and seeks to provide researchers with an Asian-centric corpus where the great majority of subjects are Asian multilinguals for whom English is an additional language, and thus provide a corpus which is complementary to the more European-focused VOICE.

First, we shall briefly introduce ACE and illustrate how it might be used by researchers. A wide range of speech events have been included in ACE: interviews; press conferences; service encounters; seminar discussions; working group discussions; workshop discussions; meetings; panels; question-and-answer sessions; conversations; etc. The transcribed speech events are categorized under five major types of setting: education (25%), leisure (10%), professional business (20%), professional organisation (35%), and professional research/science (10%). The corpus data have been tagged following the transcription conventions originally developed by the VOICE project team. These tags enable us to obtain a clear picture of the transcribed speeches (e.g., pauses, overlaps, pronunciation variations & coinages, etc.), and make ACE and VOICE comparable. Since October 2014, ACE has been officially launched online (ACE 2014). Users can browse the corpus data according to the types of setting (education, leisure, professional business, professional organization, and professional research/science.) or data collection sites (Hong Kong, Malaysia, Philippines, Vietnam, Singapore, Brunei, Japan, Mainland China, and Taiwan). A Web concordancer has been developed which allows users to search any word/phrase in ACE, and collocation information of the search word/phrase will be shown as well. Other than searching the corpus, users can also listen to the sound recording of certain ACE files, and the transcripts will be shown line by line on screen synchronously with the sound played. These functions have made it possible for researchers and

teachers/learners to explore the ACE data for various research and pedagogical purposes.

After the brief introduction of ACE, two current research studies which use data from ACE are reviewed. The first study looked at the marking or non-marking of tenses and the second investigated the communicative strategies of ELF speakers. The first study tested the hypothesis that speakers whose first languages did not mark for tense would tend to use non-marked tense forms when speaking English as a lingua franca. The hypothesis was that the substrate language would influence the speakers' English. Using a subset of ACE with speakers whose first language was a form of Malay (Malay, Bruneian Malay, Indonesian) the hypothesis was that, as Malay does not mark for tense, these speakers would thus have a tendency not to mark for tense. As will be illustrated, the hypothesis was not confirmed. In fact, these speakers with a first language of Malay were almost always found to mark for tense when the occasion was formal. Even in more informal situations, the speakers tended to mark tense more often than they did not (Kirkpatrick and Subhan 2014). This raised the question of the importance of corpora for illustrating the comparative frequency of distinctive morpho-syntactic features and the crucial significance of context and levels of formality. The results here supported recent findings of scholars such as Sharma who has argued convincingly that "the degree and distribution of a given feature must be understood in relation to the substrate before any universal claims can be made" (2009: 191).

The findings also supported those of Hall et al. (2013: 15) who, in their study of the occurrence of countable mass nouns concluded that L1 substrate influence was not high and that the countable use of mass nouns, while being widespread and attestable across different L1 backgrounds and geographical regions, was also infrequent, with a maximum occurrence rate of only 3.5%.

The second study investigated the use of communicative strategies by Asian multilinguals to see if earlier research which reported that English as a Lingua Franca is characterised by ELF speakers' adoption of specific communicative strategies to ensure successful communication and the preservation of their fellow interlocutors' face, could be supported. The editors of a review of recent trends in ELF research conclude that these trends 'evidence the supportive and cooperative nature of interactions in ELF where meaning negotiation takes place at different levels' (Archibald et al. 2011: 3). House has spoken of the 'solidarity of non-native ELF speakers' (2003: 569). Findings pointing to the cooperative nature of ELF interactions have also been reported by Firth (1996) and Meierkord (2012). Firth identified strategies such as the 'let it pass' principle, whereby speakers, instead of seeking immediate clarification when they did not understand what a speaker was saying, would let it pass, hoping, often correctly, that the meaning would become clear later. Meierkord's findings indicate that 'the conversations are characterised by their participants' desire to render the interactions normal and to achieve communicative success' (2012: 15). In a study of the communication strategies of Asian ELF speakers, Kirkpatrick (2007) identified 15 communicative strategies adopted by ELF speakers to ensure successful communication.

Once again, the study found that context was the crucial variable and that there were occasions when speakers, far from seeking to preserve the face of their fellow interlocutors, were happy to threaten it. For example, in the courtroom exchanges in the ACE data, it was found, perhaps not surprisingly, that direct, confrontational questioning and bald-on-record disagreement are common currency in these exchanges, where winning the argument supersedes the desire for interactional comity (Kirkpatrick et al. in press). The study also investigated whether there was any evidence for the so-called ASEAN way, a communicative approach based on consensus and dialogue. Some evidence for this was found and examples will be provided.

As a conclusion, based on the research findings, proposals will be made for ELT policy and pedagogy.

References

ACE. 2014. *The Asian Corpus of English*. Director: Andy Kirkpatrick; Researchers: Wang Lixun, John Patkin, Sophiann Subhan. http://corpus.ied.edu.hk/ace/ (accessed on 30 December 2014).

Archibald, A., Cogo, A. and Jenkins, J. (eds.) 2011. *Latest trends in ELF research*. Newcastle, UK: Cambridge Scholars Publishing.

Firth, A. 1996. "The discursive accomplishment of normality: On 'lingua franca' English and conversation analysis". *Journal of Pragmatics* 26: 237–259.

Hall, C. J., Schmidtke, D. and Vickers, J. 2013. Countability in world Englishes. *World Englishes* 32(1): 1-22.

House, J. 2003. English as a lingua franca: A threat to multilingualism?. *Journal of Sociolinguistics* 7(4): 556-578.

Kirkpatrick, A. 2007. "The communicative strategies of ASEAN speakers of English as a lingua franca". In D. Prescott (ed.) *English in Southeast Asia: Literacies, literatures and varieties* (pp. 121-139). Newcastle, UK: Cambridge Scholars Publishing.

Kirkpatrick, A and Subhan, S. 2014. Non-standard or new standards or errors? The use of inflectional marking for

present and past tenses in English as an Asian lingua franca. In S. Buschfeld, T. Hoffman, M. Huber and A. Kautsch et al. (eds.) *The Evolution of Englishes* (pp. 386-400). Amsterdam: John Benjamins.

Kirkpatrick, A., Walkinshaw, I and Subhan S. in press. English as a lingua franca in East and Southeast Asia: implications for diplomatic and intercultural communication. In Friedrich P (ed.) *English for Diplomatic Purposes*. Bristol: Multilingual Matters.

Meierkord, C. 2012. *Interactions across Englishes*. Cambridge, UK: Cambridge University Press.

Sharma, D. 2009. Typological diversity in new Englishes. *English World-Wide* 30(2): 170-195.

Tweet all about it: Public views on the UN's HeForShe campaign

Róisín Knight
Lancaster University
r.knight1@lancaster.ac.uk

1 Introduction

On 20th September 2014, Emma Watson, in her role as UN Women Goodwill Ambassador, gave a speech at the United Nations Headquarters through which she formally launched the UN Women's HeForShe campaign. In this speech, she argued "no country in the world can yet say they have achieved gender equality" and the word "feminist" has become "synonymous with man-hating". She therefore reached out to men and asked them to be "advocates for gender equality"- to be the "he" for "she" (UN Women 2014). The immediate media reaction suggested that opinion on the movement was divided, with some claiming that it was inspirational (Molloy 2014), others arguing that it was not in men's best interests to show support (Young 2014) and some suggesting it misrepresented feminism (McCarthy 2014). In light of this speech, and the subsequent press reaction, the purpose of this study is to investigate the public reaction to the HeForShe campaign through identifying and exploring discourses in a collection of Tweets about the campaign.

2 Methodology

Discourse analysis has influenced several methodologies previously used to explore gender and language (Baker 2014: 4). However, Baker (2014: 6) argues that much of the previous work in this area remains qualitative and based on relatively small amounts of data; he proposes that there are advantages to combining discourse analysis with corpus linguistics. Additionally, in the past, many studies carrying out an analysis of discourse in order to consider issues related to public opinion have relied primarily on more traditional forms of media (Gamson and Modigliani 1989). However, recent technological advancements have provided an opportunity to study conversations shared online. This enables the opinions of a wider range of people to captured, provides insights to views with very little time delay and can be used in diachronic studies (Potts et al. 2014; Tumasjan et al. 2010). This study combines these two approaches, carrying out a corpus-assisted discourse analysis of views expressed on Twitter about the HeForShe campaign.

#HeForShe became a trending hashtag on Twitter soon after the launch of the campaign. Twitter (n.d.)

explain that hashtags trend when their usage within tweets has recently greatly increased. It is therefore clear that many people used Twitter, and consequently the hashtag *#HeForShe*, to talk about the campaign. I collected all original tweets with this hashtag that were posted between the 20th September 2014 and 2nd October 2014 inclusive[66]. I allowed for all varieties of case, for example *#heforshe* and *#HeforShe* were also collected. These data were collected through DataSift, a platform that allows users to filter and collect online data such as tweets. In total, 190,419 tweets were collected, totaling 2,029,667 words. Another corpus, collected by Paul Baker and Tony McEnery, was also used as a reference corpus. This corpus consists of a random sample of 81,000 tweets that were posted between the 3rd February 2014 and 10th February 2014 inclusive and totals 1,110,819 words. When using Twitter data, the issue of sampling and representativeness can be particularly problematic (Baker and McEnery forthcoming). Whilst there is not space here to detail all decisions made, thought had to be given to issues such as retweets, spelling variations, the time period samples and the size of the corpora.

This study follows Burr's (1995: 48) definition of discourse; it is seen as "a set of meanings, metaphors, representations, images, stories, statements and so on that in some way together produce a particular version of events". Following the approach of Baker and McEnery (forthcoming), my analysis focused on keywords, calculated using AntConc 3.2.4w (2011). Once keywords were identified, they were grouped by hand, based on similarities in function, into categories. These categories were then analysed in more depth, through close examination of concordance lines, in order to identify discourses.

3 Findings

Through a brief analysis of concordance lines, the top 100 keywords were classified into seven categories: referential terms for those involved in the campaign; referential terms for the HeForShe campaign launch; ideology descriptors; evaluative terms; desired outcomes of the campaign; actions that will achieve the desired outcomes and inhibitors of the desired outcomes. From this, and through closer examination of concordance lines and original tweets, three different discourses became apparent; these will now be outlined.

Firstly, the discourse of the HeForShe fight was identified. 11 of the top 100 keywords were verbs signaling the actions viewed as necessary to achieve gender equality. Many of these actions are what would be expected of a soldier readying for a fight, for example *pledge, committed, stand* and *engage*. It is apparent from concordance lines that these verbs were evaluated positively and seen as ways for people to aid the campaign, yet how exactly they achieve this is unclear. For example, the word *support* occurred 9,703 times. Yet what can people do to support the campaign? Very few concrete examples are discussed. The phrase *support by* occurred only 8 times, and the phrase *support through* occurred just once. However, one function of these verbs is that they appear to be used to create a team mentality. For example *stand* occurred 9,249 times and 6,050 of instances are *stand with*.

Secondly, there is a discourse of gender through which men are frequently presented as having greater power than women. For example, some tweets shared examples of inequalities women had faced in the past, stressing the need for the campaign. However, in relation to the campaign, men were also represented as relatively more powerful. For example, Table 1 shows that *men* was more frequently positioned directly in front of keywords that signal the actions necessary to achieve gender equality, as an agent, than *women*. Please note, in total *men* occurred 13,176 times and *women* 13,530 times.

Keyword search term	Instances of 'men' as agents of the keyword	Instances of 'women' as agents of the keyword
support*	573	55
commit*	39	1
invit*	1	0
pledge	96	4
stand*	288	55
address*	5	1
engag*	20	1
perceive	0	1

Table 1. *Men* and *women* as agents

This represents men as a relatively powerful force for change, compared to women. However, there is also a minority discourse that challenges this view, as in the example below:

@KiraShip*: #heforshe is being plugged as a solidarity movement between men & women. I hope they clarify, I don't want men speaking FOR me, but WITH me.*

Finally, the discourse of Emma Watson was considered. In the top 100 keywords, 8 words were referential terms for Watson: 4 were variations of her given name *Emma Watson*; 3 were variations of

[66] I would like to thank Mark McGlashan for his guidance in collecting the corpus; for example, he provided invaluable help through writing the code required to gather the data.

Hermione Granger (the character she played within the Harry Potter films) and one was part of her official UN role of *Goodwill Ambassador*. When referring to her fictional character, the majority of tweets evaluated Watson or the campaign positively. In contrast, her official UN title was used in tweets that were largely factual as opposed to evaluative. This suggests that her fictional character was used as a way to express personal judgments.

Like *women*, Watson was rarely positioned as the agent of actions that would help the HeForShe campaign succeed. For example, the two most common referential terms were *Emma* (28,940 occurrences) and *Watson (*23,820 occurrences). Yet collectively these nouns were positioned directly in front of the keywords in Table 1, as an agent, a total of 165 times. Despite this, the vast majority of tweets evaluated her positively.

4 Conclusion

There are currently several concerns regarding the representativeness of Twitter data; however, through focusing on tweets about HeForShe, it was possible to gain an understanding of how it was perceived and presented. Such findings could potentially have real impact, if considered by the campaign. For example, it enables them to reflect upon their marketing. Baker (2008: 257) argues that regardless of what approach is taken in studying gender and language, in order for academics' work to transform the wider society it is important that the research is linked to real-life concerns. This investigation exemplifies how Twitter data offers one opportunity to achieve this.

References

Baker, P. 2008. Sexed texts. London: Equinox Publishing Ltd.

Baker, P. 2014. Using corpora to analyze gender. London: Bloomsbury Academic.

Baker, P. and McEnery, T. Forthcoming. Who benefits when discourse gets democratised? Analysing a Twitter corpus around the British Benefits Street debate.

Burr, V. 1995. An introduction to social constructionism. London: Routledge.

Gamson, W. and Modigliani, A. 1989. Media discourse and public opinion on nuclear power: A constructionist approach. American Journal of Sociology, Vol. 95, pp. 1-37.

McCarthy, A. 2014. Sorry privileged white ladies, but Emma Watson isn't a 'game changer' for feminism. Retrieved from http://www.huffingtonpost.com/xojane-/emma-watson-feminism_b_5884246.html, Accessed 16th December 2014.

Molloy, M. 2014. Emma Watson's #HeForShe campaign inspires men to fight for gender equality. Retrieved from http://www.telegraph.co.uk/culture/film/film-news/11117901/Emma-Watsons-HeForShe-UN-campaign-inspires-men-to-fight-for-gender-equality.html, Accessed 16th December 2014.

Potts, A., Simm, W., Whittle, J. and Unger, J. 2014. Exploring 'success' in digitally augmented activism: a triangulated approach to analyzing UK activist Twitter use. Discourse, Context and Media, Vol. 6, pp. 65-76.

Tumasjan, A., Sprenger, T., Sandner, P. and Welpe, I. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1441/1852, Accessed 16th December 2014.

UN Women. 2014. Emma Watson: Gender equality is your issue too. Retrieved from http://www.unwomen.org/en/news/stories/2014/9/emma-watson-gender-equality-is-your-issue-too, Accessed 16th December 2014.

Young, C. 2014. Sorry, Emma Watson, but HeForShe Is rotten for men. Retrieved from http://time.com/3432838/emma-watson-feminism-men-women/, Accessed 16th December 2014.

Ethics considerations for corpus linguistic studies using internet resources

Ansgar Koene
University of Nottingham
ansgar.koene@nottingham.ac.uk

Svenja Adolphs
University of Nottingham
svenja.adolphs@nottingham.ac.uk

Elvira Perez
University of Nottingham
Elvira.perez@nottingham.ac.uk

Chris James Carter
University of Nottingham
psxcc@nottingham.ac.uk

Ramona Statche
University of Nottingham
Ramona.statche@nottingham.ac.uk

Claire O'Malley
University of Nottingham
Claire.omalley@nottingham.ac.uk

Tom Rodden
University of Nottingham
Tom.rodden@nottingham.ac.uk

Derek McAuley
University of Nottingham
Derek.mcauley@nottingham.ac.uk

1 Introduction

With the rising popularity of public and semi-public communication channels such as Blogs (late 1990s), Wikipedia (launched in 2001), Facebook (launched in 2004), Reddit (from 2005) and Twitter (from 2006), the Internet has become an increasingly fertile medium through which to collect substantial data sets of written language. Additional features that make online communication platforms attractive include the comparatively low effort and cost associated with data collection and the unobtrusive nature of the collection process, which can often be performed 'behind the scenes' using application programme interfaces (APIs) or web scraping techniques, depending upon the affordances of the specific type of social media studies (e.g. Twitter, Blogs). While the unobtrusive nature of the methods offers the advantage of ensuring that observed conversations are not unduly influenced by the researcher, it raises ethical concerns around issues of privacy violation, informed consent and the right to withdraw.

In this paper we will discuss some of the ethical concerns around the use of online communications data. We will start by looking at the current guidelines by the British Association for Applied Linguistics (BAAL). Next we will discuss some of the core difficulties related to identifying 'publicness' of Internet-based information. This will lead to a discussion about ethical responsibilities when dealing with 'public' online communications, and how this issue is being addressed in current corpus linguistics research.

2 BAAL guidelines

In its discussion of Internet research (section 2.9) the "Recommendations on Good Practice in Applied Linguistics" guidelines (BAAL, 2006) starts by acknowledging that it is often difficult to establish if a specific online communication should be considered to be in the private or the public domain. The distinction between private and public domains, however, has significant consequences for the nature of consent and confidentiality, and how they are subsequently addressed when dealing with the data. In this respect, the BAAL (2006) guidelines advise:

> In the case of an open-access site, where contributions are publicly archived, and informants might reasonably be expected to regard their contributions as public, individual consent may not be required. In other cases it normally would be required. (BAAL, 2006)

This guideline is often interpreted to mean that when handling online data that could be reasonably considered as public communications, it is not necessary to notify participants about the act of data collection, or the analysis and publications that are based on their data. The nature of online data collection, however, is such that unless explicitly informed, participants will otherwise have no way of knowing that their communications are being observed for research purposes. This type of data collection might, therefore, also be considered as 'covert research', for which section 2.5 of the BAAL (2006) guidelines states that:

> Observation in public places is a particularly problematic issue. If observations or recordings are made of the public at large, it is not possible to gain informed consent from everyone. *However, post-hoc consent should be negotiated if the researcher is challenged by a member of the public*. (BAAL, 2006) [emphasis added by us]

The final sentence is especially important, and problematic in the context of Internet-mediated research, since the cover nature of the data collection means that participants are effectively denied this opportunity, unless the researchers make an explicit effort to inform about their actions.

Section 2.5 of the BAAL (2006) guidelines concludes with the statement that:

> A useful criterion by which to judge the acceptability of research is to anticipate or *elicit, post hoc, the reaction of informants when they are told about the precise objectives of the study*. If anger or other strong reactions are likely or

> expressed, then such data collection is inappropriate. (BAAL, 2006) [emphasis added by us]

In the context of internet-mediated research, this implies that as a minimum requirement researchers should, at the end of the data collection period, post a message about the research on the communication platform, offering some form of 'data withdrawal' procedure for any participant who wishes to make use of it. In essence, such an approach emphasises that process of 'opting-out' rather than 'opting-in'.

3 Public – Private distinction

Distinguishing between public and private communications online is probably one of the most contentious issues when trying to implement the current guidelines on Internet-mediated research ethics. Bruckman (2002) provided the following criteria for deciding if online information should be considered as 'public', and therefore, "freely quoted and analyzed […] without consent".

- It is officially, publicly archived
- No password is required for archive access
- No site policy prohibits it
- The topic is not highly sensitive.

Unfortunately, a number of potential problems persist with this sort of criteria. In a digital era characterised by Google-caching, retweeting, 'Like' buttons and other means of information replication and proliferation, what is the true meaning of "officially, publically archived"? Furthermore, in an age of 'big data' practically every communication is automatically archived by default, with publically accessible data archives generated without the user ever needing to formulate a conscious decision about the process. Blogging software, for instance, will frequently default to a mode where past blog posts are archived in a publically accessible format.

Social media, such as Twitter, introduce further problems for the Public-Private distinction. Even though it was built as a 'public broadcast' platform which people are generally aware of, it is nevertheless often used as a means for communication within networks of friends, with little intention of broadcasting content to a wider audience. In such instances, social interaction upon Twitter might be viewed more like a private conversation in a public space rather than radio broadcasts.

4 Responsibilities when dealing with public communication

A core rationale underpinning the concept that use of data from 'public' forums, such as Twitter, does not require consent is based on the premise that users of the forum have in effect already consented when they accepted the 'term and conditions' of the forum. The current reality of Internet usage, however, is that the 'terms and conditions' policies of Internet sites are rarely read and are generally formulated in ways that are too vague and incomprehensible to constitute a means of gaining true *informed* consent (Luger, 2013).

Even if users of a public forum are comfortable with the idea that their conversations may be seen and read by people they are not aware of, this does not necessarily imply that they would also be comfortable with having a large corpus of their communications analysed with the potential of generating personality profiles that intrude further into the privacy of the individual than any of their individual messages (Kosinski, Stillwell, and Graepel, 2013). This point is particularly relevant in the current climate of social media analytics where stories of unethical behaviour for commercial or security related gain are flooding the mainstream media. It is here that academia has a responsibility to enter into the discussion of what constitutes good, ethical conduct. This may be achieved by being transparent about the goals and methods of the research and gaining true *informed* consent from participants, or at least providing them with the option to withdraw.

5 Privacy vs. The greater good

So far we have discussed the issue of Ethics of Internet-based research primarily from the perspective of 'respect for the autonomy and dignity of persons', e.g. privacy. Other important ethics considerations, concerning 'scientific value', 'social responsibility', and 'maximizing benefits and minimising harm' must also be taken into consideration (BPS, 2013). When considering and conducting research studies related to preventing of socially unacceptable behaviours, such as bullying for instance, not all parties in the conversation data-set are necessarily equal. While seeking consent from the target of the bullying behaviour would be ethically required, asking consent from those who perform the bullying may not take priority in every context.

6 Conclusion

When considering the ethics of Corpus Linguistics studies using Internet-mediated resources we conclude that the concept of a binary divide between public and private communication is fundamentally flawed. Furthermore, we argue that the idea that the 'public' nature of a communication platform provides a carte-blanch for accessing the data hosted on it is highly problematic, creating a key issue for

corpus linguists who analyse different types of discourse on the internet.

Acknowledgements

This work forms part of the CaSMa project at the University of Nottinghan, HORIZON Digital Economy Research institute, supported by ESRC grant ES/M00161X/1. For more information about the CaSMa project, see [67].

References

British Association for Applied Linguistics, 2006, *Recommendations on Good Practice in Applied Linguistics*. Available online at http://www.baal.org.uk/dox/goodpractice_full.pdf

British Psychological Society, 2013. *Ethics Guidelines for Internet-mediated Research.* INF206/1.2013. Leicester.

Bruckman, A., 2002. *Ethical Guidelines for Research Online*. http://www.cc.gatech.edu/~asb/ethics/

Luger, E., 2013. *Consent for all: Revealing the hidden complexity of terms and conditions.* Proceedings of the SIGCHI conference on Human factors in computing systems, 2687-2696.

Kosinski, M., Stillwell, D. and Graepel, T., 2013. *Private traits and attirbutes are predictable from digital record of human behavior.* PNAS 110 (15): 5802-5805.

[67] 1 http://casma.wp.horizon.ac.uk/

Conceptualization of KNOWLEDGE in the official educational discourse of the Republic of Serbia

Milena Kostic

milenakostic09@gmail.com

1 Topic and aim of research

Knowledge is an abstract concept which appears in various types of discourses. It is not a physical object and to understand it we have to use other concepts which we know on the basis of our physical experience. Metaphor is a tool which gives structure and meaning to the concept of *knowledge* by emphasizing and hiding some of its aspects (Andriessen and Van Den Boom 2009).

The aim of our research is to analyze metaphorical conceptualizations of KNOWLEDGE in the contemporary Serbian language in a specific dataset of official legislative documents of the Ministry of Education, Science and Technological Development of the Republic of Serbia in order to identify the source domains which we sometimes even unconsciously use to understand this abstract concept.

When one says that knowledge is *gained*, *acquired*, *given*, *improved* etc. barely anyone can notice anything unusual in these expressions because they are so common that no one sees them as metaphorical. However, they do contain metaphors.

Our goal is to identify metaphorical conceptualizations of KNOWLEDGE in the official discourse in the field of education in Serbia because these conceptualizations influence and potentially shape conceptualizations of the given term in teacher and student discourses and overall in the public discourse of Serbia.

2 Methods and theoretical background

Conceptual analysis and corpus analysis are the two main methods that we used in our research.

We did conceptual analysis of the lexeme *knowledge* because it the most reliable for the study of abstract terms (Dragicevic 2010). The second method was corpus analysis as the results can be empirically verified when a corpus is publicly available.

The dataset consisted of 400 sentences which contained the term KNOWLEDGE in singular and plural in all cases of the Serbian language. The examples containing the target term were extracted manually by the author from the corpus of official documents such as laws, policies, regulations and strategies about the different levels of Serbian

education system. These documents were taken from the official website of the Ministry of Education, Science and Technological Development of Serbia.[68] The analyzed dataset included all the sentences from the corpus that contained the lexeme *knowledge*.

The first phase of the research was reading the documents and marking the examples which contained the target concept. Then, the examples which contained the lexeme *knowledge* used metaphorically were excerpted and grouped with similar examples to make the so-called "small" metaphors. After that the "small" metaphors were grouped so that we could identify the conceptual metaphors. Finally, we tried to determine if there was any relation between these conceptualizations and what were its implications.

The theoretical background of the research relies on the works of Lakoff and Johnson (1980), Köveczes (2002, 2010) and Klikovac (2004) who study metaphor within cognitive linguistics. Lakoff and Johnson (1980), the most influential authors in the study of the mechanism of metaphor, define it as an understanding of one concept (target domain) which is more abstract, on the basis of another concept (source domain) which is more concrete and experientially closer to people. In language, metaphors are realized in metaphorical expressions and to reconstruct a certain conceptual metaphor one needs to analyze these expressions (Lakoff and Johnson 1980). This was our task in the research.

In cognitive linguistics the meaning of linguistic expressions equals conceptualization – "forming concepts on the basis of physical, bodily, emotional and intellectual experience" (Klikovac 2004). In cognitive-linguistic theory, in addition to the objective similarity, conceptual metaphors are grounded in different types of human experience, different correlations in human experience, structural similarities and cultural roots which two concepts can share (Köveczes 2010). Therefore, conceptual metaphors, more precisely source domains, are grounded in our perceptive, biological and cultural experience (Klikovac 2004; Köveczes 2010).

3 Results

Our contact with physical entities and our own body are a basis for a number of ontological metaphors. Events, activities, emotions, ideas and other abstract concepts are understood as entities and substance (Lakoff and Johnson 1980). When one identifies their experience as entities or substance, one can refer to it, speak about it, categorize it, group and count it and in that way understand it (Lakoff and Johnson 1980). People have a need to project boundaries to a physical phenomenon in order to define it as we ourselves are defined and as entities have surfaces and boundaries (Lakoff and Johnson 1980).

Most generally, KNOWLEDGE is conceptualized as an ENTITY in our dataset and in a smaller number of examples it is conceptualized as substance. However, conceptualizing abstract concepts such as KNOWLEDGE as entities or substance does not allow us to fully understand them, but ontological metaphors can be further elaborated which is shown in Figure 1 and Figure 2.

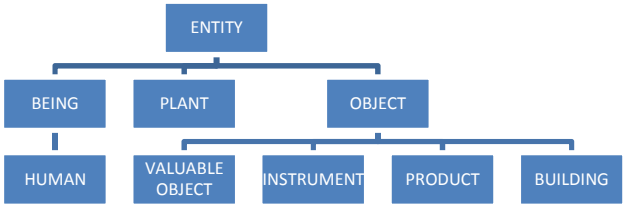


Figure 5. KNOWLEDGE IS ENTITY

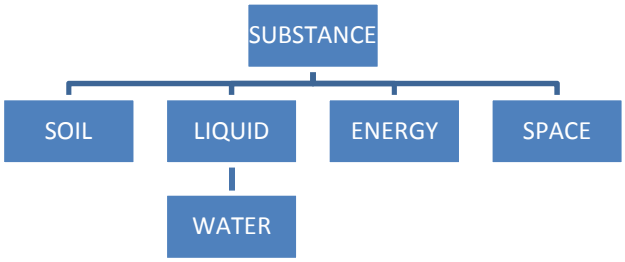


Figure 6. KNOWLEDGE IS SUBSTANCE

The basic distinction is made between metaphors KNOWLEDGE IS ENTITY (BEING, PLANT, OBJECT) (Figure 1) and KNOWLEDGE IS SUBSTANCE (SOIL, LIQUID) (Figure 2). KNOWLEDGE IS ENERGY and KNOWLEDGE IS SPACE (Figure 2) metaphors are not concretization of the conceptual metaphor KNOWLEDGE IS SUBSTANCE but they are related to it in the sense that they do not have firm boundaries as neither does substance. That is why we put them in this category and their shapes in Figure 2 are filled in with lines.

KNOWLEDGE IS ENTITY metaphor is a general ontological metaphor which is the basis for metaphors KNOWLEDGE IS BEING, PLANT and OBJECT. KNOWLEDGE IS BEING is concretized in a metaphor KNOWLEDGE IS HUMAN whose examples make 3% of our dataset. The implication of this metaphor could be that it gives the most

[68] The complete list of documents can be found on the following link: http://www.mpn.gov.rs/dokumenta-i-propisi.

active role to knowledge because it possesses the characteristics of human beings. Similarly, KNOWLEDGE IS PLANT metaphor implies dynamics of knowledge which in specific examples refers to its ability to *grow* and to *develop*. These examples make 4% of our dataset.

The results of the analysis show that KNOWLEDGE is predominantly conceptualized as an OBJECT. Examples of this metaphor make 69% of our dataset. KNOWLEDGE IS OBJECT is concretized in metaphors KNOWLEDGE IS VALUABLE OBJECT, INSTRUMENT, PRODUCT and BUILDING. Although these examples may imply the static role of KNOWLEDGE, we cannot isolate it as a general characteristic of the analyzed examples within the metaphor KNOWLEDGE IS OBJECT due to the co-text and context in which the examples occur. KNOWLEDGE is also conceptualized as an INSTRUMENT which implies its active role in undertaking certain tasks or activities.

Examples of KNOWLEDGE IS SUBSTANCE metaphor (Figure 2) make 13% of our dataset. The metaphor KNOWLEDGE IS SPACE (11% of the dataset) is connected with it because neither space nor substance has boundaries. Finally, there is only one example of the metaphor KNOWLEDGE IS ENERGY in our dataset. It is not easy to find the connection between substance and energy but they do have one characteristic in common and that is accumulation. Both energy and substance can be accumulated.

4 Conclusion

The potential importance of the outlined conceptualizations of KNOWLEDGE could be in the fact that the analyzed dataset is the basis or the starting point for the creation of discourse of education. Therefore, one can expect that these conceptualizations of KNOWLEDGE will influence the conceptualization of this term in a teaching, teacher and student discourse. Furthermore, one could analyze the conceptualizations of TEACHER and STUDENT in the same dataset to determine if these conceptualizations are connected with or conditioned by the conceptualizations of KNOWLEDGE. In that way the roles of agents, topics and aims of education could be defined and the ideologies interwoven in the education policies and strategies in the Republic of Serbia could be identified.

KNOWLEDGE is not a concept with a clearly defined structure. Whatever the structure it gets, it gets it via metaphor (Andriessen 2006). However, there are some limitations we should be aware of. In different contexts, different characteristics of KNOWLEDGE will be highlighted by metaphor. That is why we should try to identify as many conceptualizations of this term as possible in order to fully grasp it. We hope that the results of this research make a small contribution in that sense.

References

Andriessen, D. 2006. On the metaphorical nature of intellectual capital: a textual analysis, *Journal of Intellectual Capital*, 7(1), 93 –110.

Andriessen, D. 2008. Stuff or love? How metaphors direct our efforts to manage knowledge in organizations, *Knowledge Management Research & Practice,* 6, 5–12.

Andriessen, D. and Van Den Boom, M. 2009. In Search of Alternative Metaphors for Knowledge; Inspiration from Symbolism. *Electronic Journal of Knowledge Management*, 7(4), 397 – 404.

Bratianu, C. and Andriessen, D. 2008. Knowledge as Energy: A Metaphorical Analysis. *9th European Conference on Knowledge Management,* Southampton Solent University, Southampton, UK.

Dragićević, R. 2010. *Leksikologija srpskog jezika*. Beograd: Zavod za udžbenike.

Charteris-Black, J. 2004. *Corpus Approaches to Critical Metaphor Analysis*. Palgrave Macmillan.

Johnson, M. 1987. *The Body in the Mind. The Bodily Basis of Meaning, Imagination, and Reason*, Chicago: The University of Chicago Press.

Kalra, M.B. and Baveja, B. 2012. Teacher Thinking about Knowledge, Learning and Learners: A Metaphor Analysis. *Social and Behavioral Sciences* 55, 317 – 326.

Klikovac, D. 2004. *Metafore u mišljenju i jeziku,* Beograd: XX vek.

Klikovac, D. 2006. *Semantika predloga – Studija iz kognitivne lingvistike*, Beograd: Filološki fakultet (2. izdanje).

Klikovac, D. 2008. *Jezik i moć,* Beograd: XX vek.

Köveczes, Z. 2002. *Metaphor. A Practical Introduction*. Oxford: Oxford University Press.

Köveczes, Z. 2010. *Metaphor: A Practical Introduction,* Oxford: OUP (2nd ed.).

Lakoff, G. and Johnson, M. 1980. *Metaphors We Live By,* Chicago: University of Chicago Press.

Lakoff, G. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: University of Chicago Press.

Frequency and recency effects in German morphological change-in-progress

Anne Krause
University of Freiburg
anne.krause@frequenz.uni-freiburg.de

1 Introduction

Joan Bybee has dedicated a lot of her work to frequency effects in language, such as the "Conserving Effect" (Bybee and Thompson 1997: 380) of word frequency on analogical change. Already in her first explanation of this effect, she hinted at the fact that it may affect also "modern leveling", i.e. changes-in-progress or linguistic variation:

> One case I have investigated involves the six verbs *creep*, *keep*, *leap*, *leave*, *sleep*, and *weep*, all of which have a past form with a lax vowel [...]. Of these verbs, three, *creep*, *leap*, and *weep*, all may have, at least marginally, a past forms [sic] with a tense vowel, *creeped*, *leaped*, and *weeped*. The other three verbs are in no way threatened by levelling; past forms **keeped*, **leaved*, **sleeped* are clearly out of the question. [...] Again the hypothesis that less frequent forms are leveled first is supported. (Hooper 1976: 99-100)

Nevertheless, the majority of research into such frequency effects is still concerned with completed changes resulting in high-frequency irregularities and confined to a small number of languages.

The present study endeavours to overcome both gaps by examining the formation of the imperative singular in the paradigm of German strong verbs with vowel gradation like *geben* 'give'. This verb group traditionally exhibits a stem vowel change in the second and third person singular indicative and the singular imperative:

		present	
number	person	indicative	imperative
singular	1st	*gebe*	
	2nd	*gibst*	*gib*
	3rd	*gibt*	
plural	1st	*geben*	
	2nd	*gebt*	*gebt*
	3rd	*geben*	

Table 1: Conjugation table for the German verb *geben* ('give')

The imperative singular can today be found both as the variant *gib* (see Table 1) and as the analogically formed variants *geb* and *gebe* (and correspondingly for all other verbs of the paradigm). On the basis of this assumed change-in-progress, the present study aims to show that i) the validity of the Conserving Effect and other frequency effects can be confirmed for German morphology and it does not hinge on the question of linguistic variation or language change, ii) it is crucial to take into account several frequency measures and other potential explanatory factors like recency, and iii) for recent linguistic developments, the findings from corpus analyses may be enriched by those from experimental testing.

2 Data and method

German offers three forms of the imperative mood (formal plural, informal plural, and singular) of which the singular variant is restricted largely to discourse between familiar conversation partners and thus sparsely documented in corpus data, both spoken and written. However, "walkthroughs" (video game guides, i.e. step-by-step instructions by gamers for gamers) contain a high number of these forms. Therefore, a corpus containing such texts was compiled from a website for the present investigation (ca. 7m words). The search in this corpus for occurrences of all three imperative singular variants yielded a dataset of 1,939 observations.

All these observations were annotated for verb lemma token frequency, verb stem token frequency, intraparadigmatic type frequency of the e-stem, and the relative intraparadigmatic token frequency of the i-stem and the e-stem. These measures were taken from five corpus-based frequency lists or dictionaries of German which also served as the data basis for testing other hypotheses related to frequency.

Lastly, a close reading of a sample of the corpus texts strongly suggested a recency effect in the imperative formation which resulted in the annotation of the observations for previous occurrences of imperative singular forms in a predefined window of context to the left of the search hit. Mixed-effects logistic regressions were fitted to the data, testing the influence of the defined variables on the stem vowel (*gib* vs. *geb*) and suffixation (*geb* vs. *gebe*) of the imperative singular instances in the dataset.

In a subsequent experimental study, processing latencies and recall accuracies of the different imperative singular forms are treated as a token of speakers' acceptance and thus suitability for use of the analogical variants.

3 Type frequency effects

The fact that the imperative singular forms of strong verbs with vowel gradation are being replaced by

variants formed in analogy to weak paradigms could be attributed to the latter's type frequency: "the higher the type frequency the greater the productivity or likelihood that a construction will be extended to new items" (Bybee 2010: 67). In order to test this assumption, the type frequency list offered by the Institute of German Language (IDS Mannheim, DeReWo 2012) was reduced to include only verbs, which were then annotated for the paradigm to which they belong, such as the weak, the irregular, and the different strong inflectional paradigms.

The frequency distribution of verbs in the list shows that the weak paradigm applies to the majority of verb types and thus is by far the most productive of all verb conjugation patterns, whereas the paradigm of strong verbs with vowel gradation is found among the lower frequent ones. It is thus not surprising that verbs from this paradigm should fall prey to analogical levelling in direction of the weak paradigm, i.e. inflection without stem vowel change.

4 Token frequency effects

Two rather different token frequency effects which support this type frequency effect were found in the analysis of the frequency counts extracted from the corpus Wortschatz Universität Leipzig (1998-2014). On the one hand, it was observed that of the three potential candidates for analogical levelling, i.e. the 2nd and 3rd person singular and the imperative singular, the latter has the lowest relative token frequency within the paradigm of the majority of verbs (exceptions being phraseological uses). This leads to a weak mental representation of the form and makes it more prone to analogical levelling towards the weak paradigm than the 2nd and 3rd person singular indicative. On the other hand, the aggregated relative token frequency of forms with an e-stem within the verb paradigms is higher than that of the i-stem forms. This effect thus reinforces the type frequency effect explained above, viz. imperative formation without a stem vowel change to -i-.

Perhaps most importantly, the statistical analysis of the walkthrough corpus data shows that verb token frequency has a significant impact on a particular verb's imperative singular form being subject to change. More precisely, as shown in previous research, lower frequency verbs show a preference for the analogically formed (e-stem) variant, whereas high frequency verbs behave more conservatively by occurring predominantly with the established (i-stem) variant (see Figure 1).

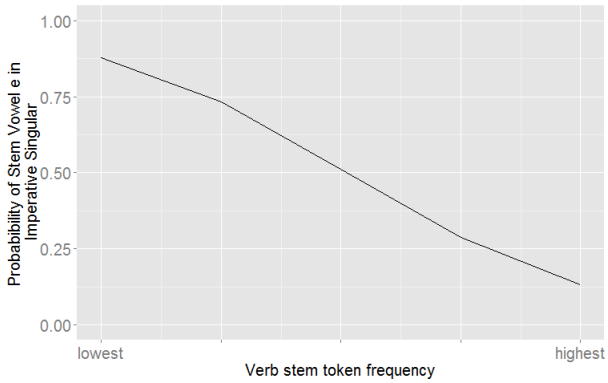


Figure 1: Conserving effect of verb stem token frequency

Note that the differentiation between a verb's lemma token frequency and stem token frequency turned out crucial to the analysis, as in the present study it was the latter measure which proved to be significant instead of the former, which is usually annotated for.

5 Recency effects

As mentioned, recency has been taken into account as an explanatory factor of the current distribution of the imperative singular forms of strong verbs with vowel gradation. It turned out to have a significant effect in the regression models, although its influence on stem vowel choice is indirect only: The suffixation of a previous imperative singular increases the probability of a suffixation in the target imperative singular forms (see Figure 2); this effect is stronger for previous imperative singulars of strong verbs with vowel gradation ("strong", e.g. *nehme* 'take') than for forms from other verb paradigms ("other", e.g. *laufe* 'run'). The suffixed i-stem form is not attested; thus, the stem vowel is adjusted to -e-.

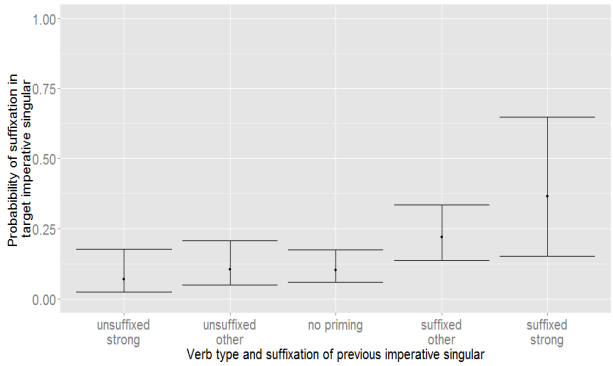


Figure 2: Recency effect of previous occurrences of the imperative singular

6 Variation or change-in-progress?

On the basis of the results of the walkthrough corpus analysis, the phenomenon under investigation is best

referred to as a "change-in-progress". A slight increase in the use of the analogical imperative singular variants is found, albeit interrupted by a strong temporary decrease, an unfortunate artefact of the data basis. The above-mentioned experimental study of the processing and perception of the three imperative singular variants involves two different age groups of participants. This apparent-time design serves to show whether the analogical imperative variants have already grown to be more accepted by younger speakers than by older participants, evidenced by lower processing latencies and higher recall accuracy rates.

7 Conclusions and outlook

Even though analogical levelling in the imperative singular of German strong verbs with vowel gradation currently is "only" in a state of change-in-progress, the frequency effects identified in the present study are robust. Thus, the Conserving effect of high token frequency and other frequency effects known from completed changes (cf. hypothesis i) have been confirmed in the present analysis.

Likewise, verb stem token frequency and recency have proved to be invaluable explanatory factors in the present model of morphological change-in-progress (cf. hypothesis ii). For the development under investigation in the present study, the analysis of corpus data could not conclusively solve one fundamental question, i.e. "variation or change?"; for such "modern leveling" phenomena experimental data can be a great asset.

References

Bybee, J. and Thompson, S. 1997. "Three frequency effects in syntax." *Proceedings of the Twenty-Third Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Pragmatics and Grammatical Structure*, 378-88.

Bybee, J. 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Hooper, J.B. 1976. "Word frequency in lexical diffusion and the source of morpho-phonological change. In W. Christie (ed.) *Current progress in historical linguistics*. Amsterdam: North Holland, 96-105.

Wortschatz-Portal Universität Leipzig. 1998-2014. <http://wortschatz.uni-leipzig.de>.

DeReWo <http://www.ids-mannheim.de/derewo>, © Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Deutschland, 2013.

Evaluating inter-rater reliability for hierarchical error annotation in learner corpora

Andrey Kutuzov
National Research University Higher School of Economics
akutuzov@hse.ru

Elizaveta Kuzmenko
National Research University Higher School of Economics
lizaku77@gmail.com

Olga Vinogradova
National Research University Higher School of Economics
olgavinogr@gmail.com

Learner corpora are mainly useful when error-annotated. However, human annotation is subject to influence of various factors. The present research describes our experiment in evaluating inter-rater hierarchical annotation agreement in one specific learner corpus. The main problem we are trying to solve is how to take into account distances between categories from different levels in our hierarchy, so that it is possible to compute partial agreement.

The corpus in question is the Russian Error-Annotated Learner English Corpus (further REALEC [69]). It comprises nearly 800 pieces of students' writing (225 thousand word tokens). Our students are mostly native Russian speakers, and they write essays in English in their course of general English. Teachers mark the essays and annotate them according to the error classification scheme (Kuzmenko and Kutuzov 2014). More than 10 thousand errors have already been annotated manually.

REALEC error annotation scheme consists of 4 layers: error type, error cause, linguistic 'damage' caused by the error and the impact of the error on general understanding of the text. The first layer of the annotation scheme in its turn consists of 151 categories organized into a tree-like structure. Annotators choose a specific tag for the error they have spotted, or apply one of the general categories in accordance with the instructions provided.

In our inter-rater reliability experiment, 30 student essays (7000 word tokens total) were chosen for this task. An experienced ESL instructor outlined error spans without marking exact error categories (520 spans in total). After that, 8 annotators were asked to assign error categories to these error spans using REALEC annotation scheme. All of them received identical guidelines. They could change the area of the error, or leave the marked span unannotated if

[69] http://realec.org

they didn't see any error in it. As a result, each annotator produced a list of error categories assigned to pre-determined error spans.

NLP community has several established means to calculate inter-rater agreement; the one most widely used is Krippendorff's alpha (further KA) (Krippendorff 2012); see (Passonneau 1997) for explanation on why precision and recall metrics are not feasible for this task. This metrics is used in our experiment. Error spans where annotation was incomplete (one or more annotators did not classify the error) were excluded so as not to complicate the data with additional variance factors. We also excluded as an outlier one particular annotator, who did not provide annotations for almost half of the predetermined error spans. This left us with only fully-annotated spans containing approximately two thirds of all the initial annotations, 2128 error category assignments in total.

First, we calculated KA for error categories from the upper level of REALEC annotation scheme. All types of grammar errors were treated as one macro-category **Grammar**; the same was done for discourse, vocabulary, etc. All macro-categories were considered to be equally 'distant' from each other. KA value was found to be 0.47 - not a very strong agreement, but still a decent one for linguistic annotation (Craggs and McGee Wood 2004).

However, to calculate inter-rater reliability for lower-level classes of our annotation scheme, we had to take into account their relative proximity to each other in the classification hierarchy. Krippendorff's alpha allows using weighted coefficients, that is to choose specific distances (levels of disagreement) between various categories. However, manually specifying distances between all possible pairs of categories is a rather time-consuming task.

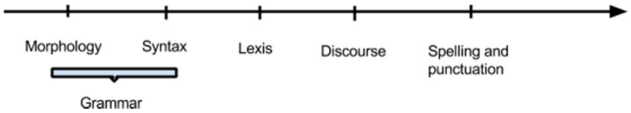


Figure 1. Continuum of annotation categories

That is why we attempted to choose another strategy and instead transformed our nominal scale into an interval one, so that KA could consider the level of disagreement (e.g., grammar errors differ one from another, but they are even more different from discourse errors, etc).

Thus, we semi-automatically assigned digital representations, or 'coefficients', to our error categories, so that tags belonging to closely related categories were assigned close values. As our scheme included five macro-categories, we placed them on the continuum of language representation levels in the way shown in the Fig. 1. Within each macro-category we assigned specific digital representations to subcategories. For example, the morphological part of macro-category **Grammar** is further divided into parts of speech to which erroneous words belong and, therefore, comprises second-level subcategories of **Verb**, **Noun**, etc. These tags are assigned different digital representations ("1", "4", "7", etc), whereas tags deeper down the hierarchy are assigned the same values as the upper ones. We also made 'gaps' 50 points wide between macro-categories.

In the next weighted annotation scheme, we went down to the third-level subcategories (for example, the ones within Grammar-Verb category: **Tense**, **Voice**, **Modals**, etc). It employed the same principles as the previous one.Thus, we were able to compute KA for interval data as if annotators had assigned interval digital values, not nominal tags. As a result, we got KA = 0.57 for the second level annotation (tags like **Noun**, **Verb**, **Word choice**, **Tautology**, etc), even higher than at the upper level. The third level annotation had agreement rate equal to 0.55. Computing KA for the second and the third annotation levels as nominal categories (binary distance) gave only 0.5 and 0.4 correspondingly.

It should be noted that when measuring 'raw' nominal KA for our data (annotations 'per se'), we got agreement value as low as 0.34. Thus, using 'digital weighing' and compressing of nominal categories allowed us to take into account more precise relationships between them. At the same time, this method leads to less clear interpretation and worse comparability of final results (Artstein and Poesio 2007).

To further prove the rationale behind the scheme in which we assigned digital representations to nominal categories, we compared it to a random 'baseline' scheme. All representations for second-level annotations assigned manually were 5 times randomly shuffled across all categories. Then we applied these quasi-schemes to annotation data and computed KA with interval distance for each case. Average agreement across the second-level tables was 0.42, significantly less than the result for manual distribution of digital representations (0.57). The result for the third-level annotations scheme was again 0.32, still lower than that for the corresponding manual scheme (0.55). This gives us ground to suppose that digital representations of annotations manage to grasp annotation inter-relations not found by simple binary distance metrics. However, this method has to be further refined and evaluated, and optimal values for the representations have to be found and proved, as currently they are based on the authors' linguistic intuition. Formulas introduced in (Geertzen and Bunt 2006) to determine relative distance between

tags in hierarchical classification can be possibly used.

Description of situation	% among cases of inconsistency	Example and solution	Course of action to be taken
Mistakes made by the annotators	33%	*…take from life all* > *…take everything from life* Tags **Nouns + Standard word order** in no way clarify the nature of the mistake made by the author.	Improving guidelines and training annotators
The same correction with different tagging	32%	*twice lucky* > *twice as* Tags **Absence of certain components in a collocation** or **Numerical comparison** don't exclude each other, both have to be assigned.	Adding guidelines on whether to apply one or both of the tags that rightfully describe the error
Multiple tags from mutually exclusive areas	15%	*…take from life all* > *...take everything from life*. Tags **Choice of a lexical item** or **Choice of determiners** together with the syntactical tag **Standard word order**. The problem was to choose between the right and the wrong determiners, so the second suggestion is better.	Training annotators to decide on which tags rightfully describe an error in difficult cases
Several corrections similarly close to the original text	13%	*Particularly distinguished New Brunswick showing…* > *New Brunswick is particularly marked off showing…* with tags **Choice among synonyms + Absence of a component in clause or sentence + Standard word order** or *New Brunswick is particularly distinguished showing…* with **Choice among synonyms + Voice form + Standard word order** or *Particularly distinguished was New Brunswick showing…* with **Absence of a component in clause or sentence + Emphatic shift**.	If the approaches suggested are equal in their proximity to the author's intention, any of them can be applied.

Table 1. Typical cases of annotations' inconsistency

The comparison of the reasons underlying cases of inconsistency in annotators' solutions was carried out manually. Table 1 lists the findings.

It is known that "the guidelines on how to annotate particular pieces of text can be elaborated almost *ad infinitum*." (Leech 2005). It is impossible to enumerate all the specific subtleties. Nevertheless, the results of this experiment will provide a solid ground for improving the existing REALEC annotation guidelines and annotators competence.

Thus, we proposed a method to employ Krippendorff's alpha in calculating inter-rater reliability in hierarchical annotation schemes, using hand-crafted digital representations of nominal categories. Also, we performed an analysis of problematic cases in annotators' inconsistency. In future we plan to optimize assignment of digital representations and support it with a straightforward algorithm and to assess the similarity of corrected variants suggested by annotators.

References

Artstein, R., and Poesio, M. "Inter-coder agreement for computational linguistics." *Computational Linguistics* 34.4 (2008): 555-596.

Craggs, R., and Wood, M. "A categorical annotation scheme for emotion in the linguistic content of dialogue." *Affective dialogue systems*. Springer Berlin Heidelberg, 2004. 89-100.

Geertzen, J., and Bunt, H. "Measuring annotator agreement in a complex hierarchical dialogue act annotation scheme." *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics, 2006.

Krippendorff, K. "Content analysis: An introduction to its methodology". Sage, 2012.

Kuzmenko, E, and Kutuzov, A. "Russian Error-Annotated Learner English Corpus: a Tool for Computer-Assisted Language Learning." *NEALT Proceedings Series Vol. 22*: 87, 2014

Leech, G. "Adding linguistic annotation". In: *Developing linguistic corpora : a guide to good practice*. Oxbow Books, Oxford, pp. 17-29, 2005

Passonneau, R. J. "Applying reliability metrics to co-reference annotation." *arXiv preprint cmp-lg/9706011* (1997).

Etymological origins of derivational affixes in spoken English

Jacqueline Laws
University of Reading
j.v.laws@reading.ac.uk

Chris Ryder
University of Reading
c.s.ryder@reading.ac.uk

1 Introduction

Complex words are coined to fill lexical gaps in the language through the addition of word-initial and word-final derivational morphemes. From an etymological perspective, derivational morphemes fall into two major classes: those derived from Germanic roots (neutral), e.g., *-ful*, and *-less*, which can be added to free bases without any change in stress; and Latinate forms (non-neutral), e.g., *-ity*, and *-ic*, which are mostly added to bound bases and often create a word stress shift, e.g., as seen in the transition from *átom* to *atómic*. Combining forms, such as *psych-ology* and *micro-cosm*, are derived from Greek and Latin lexemes.

Various authoritative sources provide useful etymological information relating to selections of prefixes and suffixes, the most notable of these being Marchand (1969). A more recent in-depth analysis of morphological etymology has been provided by Minkova & Stockwell (2009), but, in addition, Dixon (2014) has estimated that, based on a selection of around 200 affixes in English, the relative contributions of language sources are as follows:

	Prefixes (90)	Suffixes (110)
Greek	27%	10%
Romance	42%	44%
Germanic	31%	43%
Other	0%	3%

Table 1: Genetic origin of affixes (adapted from Dixon (2014:34))

Estimates of the number of derivational affixes in English vary widely. This is unsurprising given that the etymological journey of many words in present-day English presents the morphologist with a daunting challenge when attempting to arrive at a definitive answer. This exercise is thwarted for a variety of reasons. To provide just a couple of examples, some affixes are very rare and are hardy recognizable as affixes anymore in English, e.g., the Germanic and Romance nominalizing suffix *–t* that has attached to certain verbs (*give* → *gift*; *complain* → *complaint*), reported by Dixon (2014); and the fact that linguists have defined prefixes/suffixes and combining forms according to different criteria (Lehrer 1998; Prćić 2005, 2008; Stein 2007; Bauer et al. 2013). The current study employed the comprehensive classification system proposed by Stein (2007), which contains 554 word-initial and 281 word-final derivational morphemes in English.

The research reported here examined the distributional characteristics of prefixal and suffixal derivational morphemes in spoken English as a function of their language of origin. The aim of the analysis was to determine how closely the estimates provided by Dixon (Table 1) apply to affix frequency in spoken language.

2 Methodology

A corpus of complex words extracted from the spoken component of the British National Corpus (BNC) was complied. The complex words were identified by searching for all instances of the 835 derivational morphemes documented by Stein (2007). Combining forms were excluded from the current analysis owing to their predominantly classical origins and, since they tend to refer to specialised scientific genres, they are unlikely to occur in spoken language. Thus, having ascertained the etymological roots of Stein's 177 prefixes and 163 suffixes, the study focused on the occurrence (or not) of these affixes in the spoken corpus.

The version of the BNC used to compile the corpus of complex words was Davies (2012); the complete data sets from the two main sources (Demographically Sampled and Context-Governed) were combined. The grammar tagger employed was CLAWS5; Part of Speech (PoS) information was recorded for each complex word identified, and where the assigned PoS was ambiguous, the context of the item was checked in order to confirm the word class. All allomorphic variants of the master derivational morpheme set of 177 prefixes and 163 suffixes were included, for example, the prefix *in-*, as in *inappropriate*, has the three variant forms *il-*legal, *im-*proper and *ir-*rational, depending on the initial letter of the stem. This set of allomorphs constituted one prefix type.

The etymology of each complex word type was checked against the Oxford English Dictionary (OED online) to ensure that it was a true complex word containing a root and affix. The complete set of inflectional morphemes in English were added to suffix search strings to ensure that all possible word forms for each complex word were captured. Frequency values of these inflected forms were recorded and a total frequency value was assigned to each complex word type. Full details of the corpus compilation procedure can be found in Laws and Ryder (2014).

The language of origin of each affix was checked in the Oxford English Dictionary (OED online) and

classified according to the etymological groupings employed by Dixon (2014): Greek, Latin (including Romance languages, such as Old French and Italian), Germanic (including Modern, Middle and Old English, Old High German, Old Norse, Friesian, Old Teutonic and Gothic) and Other (including Russian, Sanskrit, Old Persian, Semitic, Indo-Iranian, Hebrew, Malay and Chinese), although these labels have been slightly adapted.

3 Results

The affix searches of the 10 million tokens in the spoken component of the BNC resulted in a corpus of 1,008,280 complex words, of which 986,440 contained the target prefixes and suffixes specified for this investigation, i.e., the excluded combining forms only constituted 2% of the tokens in the whole corpus of complex words.

	Prefixes	Suffixes	Totals
Target set	177	163	340
Observed set	96	141	237

Table 2: Target and Observed Affix Frequencies

Table 2 illustrates that the target set of affixes analyzed contained a few more prefixes (177: 52%) than suffixes (163: 48%), which is already an unexpected finding, given the well-established phenomenon that English is one of the languages that displays a preference for suffixation over prefixation (Greenberg, 1966). Of this target set, 96 prefixes and 141 suffixes occurred in the complex word corpus indicating that, as far as usage in spoken English is concerned, prefixes were less well-represented (41%) than suffixes (59%), as would be predicted by the suffixation preference. Moreover, a far greater number of the target suffixes (87%) occurred in the corpus than was the case for the target prefixes (39%).

Analysis of the results by language of origin revealed that the majority of the prefixes that were underused in the spoken corpus were from Greek. These prefixes tended to be more appropriate for formal, academic or scientific contexts; examples include *meta-*, as in *metaphysical*, *acousto-*, as in *acousto-electric* and *litho-* as in *lithograph*, so it is not surprising that they did not occur in spoken language. By contrast, with respect to suffixes, the observed frequencies were very close to the target frequencies for all language sources.

The distribution pattern of language source across affix types was very similar to that predicted by Dixon (2014), as presented in Table 1: the largest group of affixes came from Latin origins and affixes of Greek origin are considerably under-represented in the spoken data. However, the representation of Germanic sources was found to be less frequent than was predicted by Dixon's estimates, particularly with respect to suffixes.

4 Conclusions

The suffixation preference for English was observed, even though more prefixes than suffixes were analysed: the suffix target set was better represented in the spoken corpus than the prefix target set. This study confirmed, with a larger number of affixes, the general distributional patterns of affix etymology proposed by Dixon (2014), except that, surprisingly, a greater proportion of Latinate affixes were observed in the spoken corpus of complex words than predicted.

References

Bauer, L., Lieber, R. & I. Plag. 2013. *The Oxford Reference Guide to English Morphology*. Oxford: Oxford University Press.

Davies, M. (2012). *BYU-BNC* [Based on the British National Corpus from Oxford University Press]. Available at <http://corpus.byu.edu/bnc>.

Dixon, R.M.W. 2014. *Making new words: morphological derivation in English*. Oxford: Oxford University Press.

Greenberg, J. H. (1966). Some universal of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (ed), *Universals of Language*, 2nd edition, Cambridge, Mass: MIT Press.

Laws, J. V. & C. Ryder. 2014. Getting the measure of derivational morphology in adult speech: A corpus analysis using *MorphoQuantics*. University of Reading: *Language Studies Working Papers*, 6. pp. 3-17.

Lehrer, A. (1998). Scapes, holics, and thons: the semantics of English combining forms. *American Speech 73*, 3-28.

Marchand, H. 1969. *The categories and types of present-day English word-formation*. Second edition. München: C.H. Beck.

Minkova, D. and Stockwell, R. 2009. *English words: history and structure*. Second edition. Cambridge: Cambridge University Press.

Prćić, T. (2005). Prefixes vs initial combining forms in English: a lexicographic perspective. *International Journal of Lexicography 18*, 313-334.

Prćić, T. (2008). Suffixes vs final combining forms in English: a lexicographic perspective. *International Journal of Lexicography 21*, 1-22.

Stein, G. 2007. *A dictionary of English affixes: their function and meaning*. Munich: Lincom Europa.

Doing the naughty or having it done to you: agent roles in erotic writing

Alon Lischinsky
Oxford Brookes
alon@lischinsky.net

Once imprisoned in "secret museums" and hidden from the view of the general public (Kendrick, 1997), pornography has become an increasingly visible and important part of cultural life over the past 50 years. Visual and written representations of sexual activity, formally banned as obscene in most Western countries since the mid-19th century, entered the mass market in the 1960s and their circulation grew very significantly with the proliferation of special-interest magazines in the 1970s, the launch of home video systems in the 1980s, and the commercial internet in the 1990s (Hardy, 2009). Online pornography in particular has disrupted the various barriers historically erected to regulate access to erotic materials, giving rise to heated discussions about acceptable forms of sexual knowledge, sexual freedom and sexual representations (Atwood, 2010).

This newfound visibility has resulted in a dramatic increase in the amount and range of scholarly work on porn. While academic research on the subject until the 1990s tended to focus on alleged undesirable effects of porn consumption —such as undermining traditional values of monogamy and emotional attachment (Zillmann, 1986), or enticing men to sexual violence against women (Mackinnon 1989)— current work adopts a much more nuanced view of the various forms in which porn is consumed, of its psychological and social functions, and of its aesthetic and cultural significance (Wicke 1991). But while this scholarship has led to increasing awareness of the various forms of pornographic expression, it has largely focused on the visual genres of photography and film, and exploration of the language of contemporary porn remains limited and uneven (Wicke 1991:75).

Certain corners of this broad field have received a certain degree of attention; analyses of erotic writing belonging to the canonical genres of high literature are not rare, and there has been considerable interest in specific forms of amateur erotica, such as slash fiction, a genre of fan writing that introduces romantic or erotic elements between fictional characters that are not so paired in the original work (e.g., Dhaenens *et al.*, 2008). However, these strands of research have rarely concerned themselves with the linguistic and semiotic substance of erotic writing. The field of "pornolinguistics" imagined by McCawley (Zwicky *et al.*, 1992) remains almost entirely unpopulated, and few systematic empirical descriptions of the language of porn are available (among the few exceptions, see Dwyer, 2007; Johnsdotter, 2011).

Such an absence is especially unfortunate because the analysis of the language used to depict sexuality and sexual activity is uniquely positioned to contribute evidence of the popular understanding of these issues. Unlike literary forms of erotica —in which issues of aesthetics, stylistics and narrative form are likely to be an important concern— and commercially-produced erotica such as that published in magazines like *Forum* —in which both obscenity laws and target market considerations contribute to shaping the content and form of the stories— amateur erotic writing provides relatively direct access to the cognitive scripts according to which the wider public conceives of sexual activity. This is not to say, of course, that this kind of narratives provides an accurate representation of the sexual lives of their authors, but rather that these narratives reflect the forms of sexual practice that their authors find exciting, desirable or alluring without gatekeeping by editors or producers.

My particular focus in this study is on the linguistic representation of sexual agency in amateur online erotic writing. Traditional conceptions of pornography have claimed that it is characterised by a male-centric perspective in which females play a predominantly passive role (e.g., Mackinnon 1989). From a functional linguistic point of view, these claims can be investigated through an analysis of transitivity structures, where the participant roles linked to specific processes can be seen as the authors' encoding of their experiential reality (Halliday 1973:134). Transitivity analysis focuses on how an author represents who undertakes an action, and who is acted upon or affected. This form of analysis has long been used in critical discourse studies to investigate which social groups or roles are represented as having power over others (Trew 1979). Carter (1997:13), for example, shows that stories in women's magazines tend to cast males as agents of transitive verbs of which females are the goal.

In this paper, I apply these notions to investigate the syntax of verbs representing sexual activity in a large corpus of online erotica. Literotica.com, one of the oldest and largest erotic fiction repositories online, was used to collect a sample of the most read 500 individual stories; items from the "How To", "Novels and Novellas" and "Reviews & Essays" categories were excluded to ensure genre uniformity. The resulting corpus comprised just under 1.5 million words. Using both standalone tools and the online Sketch Engine, the syntactic patterns in which the lemma FUCK and its synonyms were

investigated. A comparison of stories written by male and female authors was also conducted, in order to determine whether the conceptualisation of sexual agency is linked to the gender of the author, or is instead determined by the norms of the genre.

References

Attwood, F. (2010). Porn studies: from social problem to cultural practice. In F. Attwood (Ed.), *porn.com: Making Sense of Online Pornography*. Oxford: Peter Lang (pp. 1–13). New York: Peter Lang.

Carter, R. (1997). *Investigating English Discourse: Language, Literacy, Literature*. London: Routledge.

Dhaenens, F., Bauwel, S. V., & Biltereyst, D. (2008). Slashing the Fiction of Queer Theory Slash Fiction, Queer Reading, and Transgressing the Boundaries of Screen Studies, Representations, and Audiences. *Journal of Communication Inquiry*, 32(4), 335–347. doi:10.1177/0196859908321508

Dwyer, R. A. (2007). Terms of Endearment? *Power and Vocatives in BDSM Erotica* (Master's thesis). University of Edinburgh, Edinburgh.

Halliday, M. A. K. (1973). *Explorations in the functions of language*. London: Edward Arnold.

Hardy, S. (2009). The new pornographies: Representation or reality? In F. Attwood (Ed.), *Mainstreaming Sex: The Sexualization of Western Culture* (pp. 3–18). London: IB Taurus & Co.

Johnsdotter, S. (2011). "The Flow of Her Cum": On a Recent Semantic Shift in an Erotic Word. *Sexuality & Culture*, 15(2), 179–194. doi:10.1007/s12119-011-9089-y

Kendrick, W. (1997). *The Secret Museum: Pornography in Modern Culture*. Berkeley, CA: University of California Press.

MacKinnon, C. A. (1989). *Toward a feminist theory of the state*. Cambridge, MA: Harvard University Press.

Trew, T. (1979). "What the Papers Say": linguistic variation and ideological difference. In R. Fowler, B. Hodge, G. Kress, & T. Trew (Eds.), *Language and control* (pp. 117–156). London: Routledge.

Wicke, J. (1991). Through a Gaze Darkly: Pornography's Academic Market. *Transition*, (54), 68–89. doi:10.2307/2934903

Zillmann, D. (1986). Effects of Prolonged Consumption of Pornography. Paper written for the *Surgeon General's Workshop on Pornography and Public Health*. Arlington, VA. Retrieved from http://sgreports.nlm.nih.gov/NN/B/C/K/V/_/nnbckv.pdf

Zwicky, A. M., Salus, P. H., Binnick, R. I., & Vanek, A. L. (Eds.). (1992). *Studies out in Left Field: Defamatory essays presented to James D. McCawley on his 33rd or 34th birthday* (reprint of the original edition). Amsterdam: John Benjamins.

Who says what in spoken corpora?: *speaker identification* in the Spoken BNC2014

Robbie Love
Lancaster University
r.m.love@lancaster.ac.uk

Claire Dembry
Cambridge University Press
cdembry@cambridge.org

1 Introduction

The Centre for Corpus Approaches to Social Science (CASS) at Lancaster University and Cambridge University Press (CUP) are collaborating on a new corpus, the Spoken British National Corpus 2014 (Spoken BNC2014). This will be the first publicly-accessible corpus of spoken British English since the spoken component of the original British National Corpus (Leech 1993) (henceforth Spoken BNC1994).

2 Speaker identification in spoken corpus transcription

As part of the compilation of the Spoken BNC2014, we developed a new transcription scheme, the definition of which functioned not only as an essential preparatory step, but also a locus for the critical examination of certain issues in corpus construction. One of the issues which arose during this process was *speaker identification* at the transcription stage. The issue in question is the role of the transcriber in accurately identifying the speaker who produced each transcribed utterance, as opposed to the actual linguistic content of the utterance. There are two practically unavoidable deficiencies in the transcription of audio recordings: transcribers' lack of familiarity with (i) the speakers and (ii) the context in which the conversation occurred. Either or both of these could lead to inaccuracies in speaker identification

The audio recordings for the Spoken BNC2014 have been provided remotely by participants from all over the UK. They were then transcribed by a small group of freelancers at Cambridge University Press. As such, none of the transcribers were present at any of the recording sessions, and furthermore, the likelihood of any individual transcriber being personally familiar with any of the speakers in the recordings (and thus being able to recognise their voice) is effectively zero. With no memory of the interaction or familiarity with the speakers to rely upon, the transcribers had to guess the speaker of each turn as best they could, as well as transcribing the linguistic content and applying transcription

conventions throughout.

Love (2014) reports on a methodological pilot study in advance of the construction of the Spoken BNC2014, using sound recordings from an earlier project. Love's findings suggest that speaker identification appears to be largely unproblematic for recordings which contain fewer than four speakers, but that recordings with four or more speakers are increasingly likely to prove difficult for transcribers. As data collection for the corpus proper progressed, it became clear that a substantial number of recordings (approximately 20%) contain four or more speakers. We therefore decided to revisit the issue in greater detail, using recordings actually collected to form part of the Spoken BNC2014.

Speaker identification is important because it is the speaker ID codes in the corpus that allow users to carry out sociolinguistic investigations, comparing the language of speakers according to demographic metadata, such as *gender, age,* or *socio-economic status* (see for instance Baker 2014; Xiao and Tao 2007; McEnery and Xiao 2004). It has been shown that making sociolinguistic generalisations based on corpus data is something that is easy to do badly (Brezina and Meyerhoff 2014). If we were to have reason to believe that a substantial number of speaker identifications in the corpus might be inaccurate, there are further worrying implications for the reliability of existing and future studies which depend upon dividing spoken corpora according to categories of demographic metadata. This being the case, it is essential for us to attempt to estimate the likely extent of faulty speaker identification in a corpus such as the Spoken BNC2014.

3 Method

An ideal method for this investigation would be to compare transcribers' efforts at speaker identification with a 'gold standard', that is, with a set of existing transcriptions of the same recordings in which all speaker identifications are known to be correct. However, no such gold standard exists. The only way one might create one would be to submit a transcription back to the participant who made the recording, and ask them to correct the speaker identification using their personal and contextual knowledge. Even this, however, would not lead to 100% reliable speaker identification,

Thus, there was no way to compare the assignation of speaker ID codes in the Spoken BNC2014 texts to a set of 'correct answers', since no such set exists. *Accuracy* of speaker identification in the corpus is therefore impossible to directly ascertain. For this reason, we devised three investigations which, while not directly measuring speaker identification accuracy in Spoken BNC2014 transcripts, do aim to provide a very clear idea of how likely it is that the transcribers identified speakers accurately. The first two investigations were carried out by comparing multiple transcribers with each other. We provided an actual legitimate Spoken BNC2014 audio recording with six speakers to several transcribers, and compared the identifications in the resulting transcripts. The third investigation involved creating a fake gold standard transcript by recording and transcribing a conversation among one of the authors and seven other speakers, and thus guaranteeing 100% accurate speaker identification. We then gave the recording used to create the manufactured gold standard transcript to the same group of Spoken BNC2014 transcribers, and compared the speaker identifications in the resulting transcripts to the manufactured gold standard.

We assessed:

- *certainty* – the average confidence of the transcribers regarding their identifications. This was based on calculating the average proportion of turns in a sample BNC2014 transcript that were marked with definite speaker ID codes as opposed to indefinite speaker ID codes.
- *inter-rater agreement* – the extent to which multiple transcribers agree regarding the speaker identification of each individual turn in a sample Spoken BNC2014 transcript.
- *accuracy* – the extent to which the transcribers matched the speaker identification of each individual turn of a manufactured, 100% accurate transcript.

4 Results

All three investigations confirm that there is cause for concern with regards to speaker identification in spoken corpus transcription.

- *Certainty* – on average, a large majority of turns were assigned definite speaker ID codes; only in a small proportion of cases did transcribers avail themselves of the indefinite speaker ID flag. This suggests that the transcribers' confidence in their own speaker identification was high.
- *Inter-rater agreement* – inter-rater agreement regarding speaker identification was alarmingly low, especially in light of the high degree of certainty noted above.
- *Accuracy* – more transcribers failed than succeeded in replicating the speaker ID coding of the manufactured transcript to anywhere near 100%.

5 Conclusion

Our investigation shows that, while transcribers are generally willing to provide a definite speaker ID code for transcribed turns (implying high confidence of speaker identification), both inter-rater agreement and accuracy relative to a manufactured gold standard relatively low.

Therefore, while we are unable to directly assess accuracy of speaker identification in actual Spoken BNC2014 transcripts, we are very confident that a substantial number of texts in the finished corpus will contain a high proportion of inaccurate speaker identifications. This is a heretofore unconsidered and yet potentially major problem for researchers who wish to use the Spoken BNC2014 (and other spoken corpora with demographic speaker metadata) for sociolinguistic research. To ameliorate the problem, we recommend that the potential for inaccurate speaker identification is clearly and comprehensively documented, both in any corpus manual(s) and in the text-level and utterance-level metadata, such that users of the corpus have the option to exclude from any given analysis the utterances or transcripts which are most likely to have fallen victim to poor speaker identification.

Acknowledgements

We are grateful to the assistance of Samantha Owen, Laura Grimes, Imogen Dickens and Sarah Grieves at Cambridge University Press, and the freelance transcribers who worked on the corpus. The research presented in this paper was supported by the ESRC Centre for Corpus Approaches to Social Science, ESRC grant reference ES/K002155/1.

References

Baker, P. 2014. *Using corpora to analyse gender.* London: Bloomsbury.

Brezina, V., & Meyerhoff, M. 2014. Significant or random? A critical review of sociolinguistic generalisations based on large corpora. International Journal of Corpus Linguistics, 19(1), 1-28. doi:10.1075/ijcl.19.1.01bre

Leech, G. 1993. 100 million words of English. *English Today*, 9-15. doi:10.1017/S0266078400006854

Love, R. 2014. *Methodological issues in the compilation of spoken corpora: the Spoken BNC2014 pilot study*. Lancaster University: unpublished MA dissertation.

McEnery, A., & Xiao, Z. 2004. Swearing in modern British English: the case of fuck in the BNC. *Language and Literature, 13*(3), 235-268. doi: 10.1177/0963947004044873

Xiao, R., & Tao, H. 2007. "A corpus-based sociolinguistic study of amplifiers in British English.". Sociolinguistic Studies, 1 (2), 241–273.

Using OCR for faster development of historical corpora

Anke Lüdeling
Humboldt-Universität zu Berlin
anke.luedeling@rz.hu-berlin.de

Uwe Springmann
CIS, Ludwig-Maximilians-Universität München
springmann@cis.uni-muenchen.de

1 Introduction

This paper describes the first open-source, high accuracy method for digitizing early prints. With the help of optical character recognition (OCR), the expensive task of manual transcription of historical documents with their typographical peculiarities such as historical ligatures, Gothic scripts and alphabet mixtures as well as historical spellings can be automated to a large extent. This method opens up the possibility of building historical corpora on a larger scale and both faster and cheaper than before.

We will exemplify the method using the RIDGES corpus of German herbal texts ranging from 1487 to 1870 built at Humboldt-University Berlin.[70]

The construction of historical corpora is time consuming and difficult (see Claridge 2008, among many others). Depending on the age of the texts and the nature of the originals most of the work has to be done manually, and many steps require expertise for a given language stage and knowledge of typography, printing processes, and the subject matter.

The first step in building a historical corpus is always the preparation of an electronic version of the text. This paper is concerned only with this first step. For each text in the RIDGES corpus, we have about 30 pages of highly diplomatic manual transcriptions.[71] It is often assumed that it is easier to correct the OCR output for a scanned facsimile than to produce the digitized version from scratch. As more and more scans (often with some form of OCR) become available from initiatives such as Google Books or library digitization programs, most of the texts in the RIDGES corpus can be found online in some version. However, the quality of the OCR is often so bad – especially for incunabula and other early prints – that the correction time is still

[70] http://korpling.german.hu-berlin.de/ridges/index_en.html; RIDGES stands for **R**egister **i**n **D**iachronic **Ge**rman **S**cience. The corpus is deeply annotated (in a multi-layer format) and freely available under the CC-BY license. It can be downloaded in several formats as well as queried through the ANNIS search tool (Krause and Zeldes, 2014).

[71] In transcribing a text, one has to take many decisions with respect to diplomaticity. For the specific decisions in RIDGES, see the manual on the homepage.

considerable with no or only a small advantage compared to transcription from scratch. A better method would therefore be highly welcome.

2 OCR for historical documents

Traditional OCR methods have been developed and applied primarily to 20th century documents. Only recently has it become possible to recognize Gothic (Fraktur) script, in which many documents in European countries have been printed in earlier centuries. Both the proprietary ABBYY Finereader[72] as well as the open-source engines Tesseract[73] and OCRopus [74] are now able to convert Gothic documents with pre-trained models covering a variety of such scripts. However, earlier printings are still largely inaccessible to OCR methods due to a variety of peculiarities ranging from unfamiliar scripts over script and alphabet mixing (e.g. Antiqua and Fraktur typefaces, Latin and Greek characters in the same document) to page deterioration from age and usage. The oldest documents, collectively known as incunables when printed between 1450 and 1500, are deemed impossible to OCR (Rydberg-Cox 2009).

The state of the art for Latin scripts and the effect of trainability of the open-source engines are summarized in Springmann et al. (2014). Character recognition based upon recurring neural nets (RNNs) with long short-term memory (LSTM; Hochreiter and Schmidhuber 1997) have recently been incorporated into the OCRopus system and shown to give excellent results on 19th and 20th century printings (Breuel et al. 2013). This has prompted us to experiment with OCRopus (version 0.7) and train it on a range of historical printings taking the available RIDGES corpus page images and their associated diplomatic transcription as training material (ground truth). The effect of the training on about 10 to 20 pages leads to character recognition rates on unseen test pages from 95% to over 99% in all cases without the use of a dictionary or any postcorrection. This is a sizable increase from previous state-of-the-art rates of about 85% (Springmann et al. 2014, Reddy and Crane 2006).

The details of the training procedure will be published in tutorial form[75], here we report on the use of OCR for extending the corpus from its current 30 pages excerpt per book to whole books and to other books with similar typefaces.

Figure 1: Uncorrected OCR output of a 1487 printing of Gart der Gesundheit showing three remaining errors (ſpucht → ſpricht, dei → der)

3 Experiments and Results

To judge the efficiency of an OCR-based method of corpus construction, the following variables must be considered: The time it takes to train an OCR model, the time to convert page images into electronic text, including any necessary preprocessing of scanned page images, and the time to correct the remaining errors from the conversion process.

The training of an OCR model needs some diplomatically transcribed pages. Training is best organized as a bootstrapping process by which a few pages are manually transcribed, then a model is trained on these pages that will be applied to new pages, where one then only needs to correct the recognition errors, followed by further training on an extended training set. After about 10-30 pages of ground truth (depending on the distribution of rare glyphs) further gains in recognition accuracy flatten out and the training process can be stopped. Machine time for training is considerably less important than manual time for setting up the training environment including ground truth production. One can gain an additional few percentage points in recognition accuracy by not relying on the OCR engine's built-in preprocessing but doing some manual preprocessing instead. With good scans, this often leads to accuracies better than 98%. The remaining effort consists in manually correcting this output to 100% or the closest approximation thereof attainable by a human.

We conducted an experiment for the earliest and most difficult to transcribe RIDGES document, the incunable “Gart der Gesundheit” (printed in Ulm in

[72] http://finereader.abbyy.com/
[73] https://code.google.com/p/tesseract-ocr/
[74] https://github.com/tmbdev/ocropy
[75] https://code.google.com/p/cistern/

1487) by Johannes Wonnecke von Kaub[76]. The OCR model was trained on nine pages and evaluated against additional five pages. Manual preprocessing was negligible. The resulting OCR character accuracy is between 96% and 97%. Figure 1 shows some lines of the original text together with the OCR output. Text in this quality is already useful for a variety of research questions, because a high recall can be achieved with spelling-tolerant search. For higher accuracy, the OCR output needs to be corrected. Table 1 shows a comparison of the effort of manual transcription against OCR correction for two different pages each. Three different annotators with varying degrees of training on incunable transcription performed both tasks. OCR correction took about half the time as manual transcription. Inter-annotator character agreement was at 98.6% for transcription and 99.03% for correction after the annotators had corrected the original transcription of another annotator within additional 20 minutes. This result shows another benefit of the OCR approach: Starting from OCR output leaves only the remaining errors as an opportunity for annotator disagreement and consequently leads to better overall agreement and higher quality of the final text.

	manual transcription	*OCR correction*
annotator 1	40	21
annotator 2	80	28
annotator 3	60	35
average/page	30	14

Table 1: Comparison of effort (in minutes) for text production of two pages each: manual transcription from scratch and correction of OCR recognition for three annotators.

This shows that the bulk work of conversion can be done automatically by a trained OCR model while the postcorrection effort is less than 50% of the effort of manual transcription. The training itself does not need any manual intervention apart from the initial production of ground truth. The postcorrection time can be reduced further by using advanced methods of interactive postcorrection by inducing error series on a whole document which can be corrected in batch mode (Vobl et al. 2014).

4 Summary

We have shown that the new trainable OCR methods based upon LSTM RNNs can provide high character accuracy (from 95% to over 99%) for images of early prints. After the initial diplomatic transcription of 10-20 pages, training an OCR model for the recognition of the rest of the book pages leads to electronic text that may already be sufficient for many research questions. A manual postcorrection based on these OCR results is considerably less time consuming than manual transcription from scratch. If the postcorrection task is given to different annotators whose results will later be compared for further error reduction, inter-annotator agreement is higher for OCR-corrected text than for manual transcriptions. The gain of an OCR approach will be even higher if the models generalize to collections of books with the same typeface and if advanced postcorrection methods enabling batch correction of common errors are applied.

[76] See the scan at http://daten.digitale-sammlungen.de/bsb00048197.

References

Breuel, T. M. 2008. “The OCRopus open source OCR system.” *Electronic Imaging 2008*.

Breuel, T. M., Ul-Hasan, A., Al-Azawi, M. A & Shafait, F. 2013. “High-performance OCR for printed English and Fraktur using LSTM networks.” In *12th International Conference on Document Analysis and Recognition (ICDAR),* 683-687.

Claridge, C. 2008. Historical corpora. In A. Lüdeling &M. Kytö (eds.) *Corpus Linguistics. An International Handbook*, Berlin: Mouton de Gruyter, 242–259.

Hochreiter, S. and Schmidhuber, J. 1997. „Long short-term memory.“ *Neural computation*, *9(8)*, 1735-1780.

Krause, T. and Zeldes, A. 2014. “ANNIS3: A new architecture for generic corpus query and visualization”. *Digital Scholarship in the Humanities*.

Reddy, S. &Crane, G. 2006. “A document recognition system for early modern Latin.” *Chicago Colloquium on Digital Humanities and Computer Science: What Do You Do With A Million Books, Chicago, IL.*

Rydberg-Cox, J. A. 2009. “Digitizing Latin incunabula: Challenges, methods, and possibilities.” *Digital Humanities Quarterly*, 3(1).

Springmann, U., Najock, D., Morgenroth, H., Schmid, H., Gotscharek, A. & Fink, F. 2014. “OCR of historical printings of Latin texts: problems, prospects, progress.” *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage,* 71-75.

Vobl, T., Gotscharek, A., Reffle, U., Ringlstetter, C. &Schulz, K. U. 2014. “PoCoTo an open source system for efficient interactive postcorrection of OCRed historical texts.” *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage,* 57-61.

Increasing speed and consistency of phonetic transcription of spoken corpora using ASR technology

David Lukeš
Charles University, Institute of the Czech National Corpus
david.lukes@ff.cuni.cz

1 Introduction

As William Labov (e.g. Labov 1963) has amply demonstrated, phonetic variables can pattern in informative ways with respect to sociolinguistic factors. Similarly, a quantitative analysis of the connected-speech processes in a language might help us understand which kinds of sound change it is prone to undergo, and whether they are perhaps lexically constrained (lexical diffusion, see e.g. McMahon 1994). Adding phonetic transcription to sociolinguistically diverse spoken corpora should therefore be a natural choice, all the more since they tend to be smallish and many phonetic phenomena necessarily have a higher rate of recurrence than word-level phenomena.

Yet producing such a transcription is time-consuming and costly: it requires a considerable amount of manual work from human experts. By the same token, it is also error-prone and potentially inconsistent in large projects: it is hard to maintain consistency over a span of several years, in spite of stringent quality control.

The spoken corpus currently in development at the Institute of the Czech National Corpus, called ORTOFON (see Kopřivová et al. 2014), will include a phonetic layer in addition to the basic transcript. Manual work on transcribing is well under way, and we are now exploring ways of automating the process to alleviate the issues sketched out above.

2 Forced alignment

Forced alignment is an iterative technique in Automatic Speech Recognition (ASR) whereby (recording, basic transcript) pairs are time-aligned on the phone level with respect to each other (Young 2009). The whole process bootstraps, its only additional input is a pronouncing dictionary to retrieve the pronunciations of the basic transcript tokens.

Crucially, this dictionary can feature several pronunciations for any word; at each iteration of the alignment procedure, a variant is picked so as to maximize the probability of the overall alignment, i.e. with respect to the actual sound of the recording and the acoustic models of the phones as they have been estimated so far.

In the case of the ORTOFON corpus, running a forced alignment would thus enable us to:

- verify existing transcriptions for consistency
- generate a sound-aware initial phonetic transcription, which human experts would then only post-edit, thereby speeding up the process

Unlike ASR proper, this method needs to leverage a pre-existing basic transcript and pronunciation dictionary; on the other hand, it does not necessitate any pre-trained acoustic or language models, both of which are problematic when it comes to spontaneous speech data.

3 Additional benefits

A side effect of this process is a phone-level alignment of the transcript with the sound. With high-quality recordings, this will enable detailed phonetic analyses (e.g. formant extraction). With lower-quality recordings, some more robust features like duration and fundamental frequency (F0) contours can still be extracted. As acoustic correlates of prosodic features (rhythm/timing and intonation, respectively), these can in turn be used to easily add some rudimentary prosodic annotation to the corpus.

4 Problems encountered and potential

Pronunciation variants for out-of-vocabulary (OOV) words (i.e. words not yet phonetically transcribed) can either be rule-induced[77] and/or generated by a stochastic grapheme-to-phoneme converter such as Sequitur G2P (Bisani and Ney 2008). If this results in too many variants, similar ones can be collapsed, either manually or using an edit distance measure (perhaps adjusted for phonetic similarity, see e.g. Kessler 2007).

A confidence metric (cf. Jiang 2005) on the pronunciation variant that is ultimately selected would be useful in theory to highlight tokens to pay attention to while post-editing, but its practical value as tested on actual non-studio speech is problematic (Brocki et al. 2014).

As to the phone-level alignment, manual correction remains necessary if high reliability and accuracy are required (Machač and Skarnitzl 2009). Similarly, the transcription of overlapping speech will always demand more extensive manual verification in single-microphone setups.

[77] This makes sense for Czech, where the grapheme-to-phoneme correspondences are relatively straightforward. English would rather use a reference pronouncing dictionary such as CMUdict (Weide 1998) for AmE or BEEP (Robinson 1997) for BrE.

5 Current state of project

A proof-of-concept of the basic premise of this approach has been successfully tested on typical ORTOFON data using the HTK ASR toolkit (Young et al. 2009), but we are in the process of migrating our experimental setup to KALDI (Povey et al. 2011), a state-of-the-art toolkit under active development.

The pronouncing dictionary is easily derived from the set of existing ORTOFON transcriptions. An initial implementation of the rule-based grapheme-to-phoneme conversion for OOV words is complete, and we are currently looking into Sequitur G2P's abilities for stochastic modeling of the same.

Acknowledgements

This paper resulted from the implementation of the Czech National Corpus project (LM2011023) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

References

Bisani, M. and Ney, H. 2008. "Joint-sequence models for grapheme-to-phoneme conversion". *Speech Communication* 50: 434-451.

Brocki, Ł., Koržinek, D. and Marasek, K. 2014. "Challenges in Processing Real-Life Speech Corpora". Presentation at *Practical Applications of Language Corpora (PALC 2014)*. Łódź, Poland.

Jiang, H. 2005. "Confidence measures for speech recognition: A survey". *Speech Communication* 45: 455-470.

Kessler, B. 2007. "Word Similarity Metrics and Multilateral Comparison". In *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*. Prague, Czech Republic: Association for Computational Linguistics.

Kopřivová, M., Klimešová, P., Goláňová, H. and Lukeš D. 2014. "Mapping Diatopic and Diachronic Variation in Spoken Czech: the ORTOFON and DIALEKT corpora". In N. Calzolari et al. (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*. Reykjavik, Iceland: European Language Resources Association (ELRA). Available online at http://www.lrec-conf.org/proceedings/lrec2014/pdf/252_Paper.pdf

Labov, W. 1963. "The Social Motivation of a Sound Change". *Word* 19: 273-309.

Machač, P. and Skarnitzl, R. 2009. *Principles of Phonetic Segmentation*. Prague, Czech Republic: Nakladatelství Epocha.

McMahon, A.M.S. 1994. *Understanding Language Change*. Cambridge: CUP.

Povey, D. et al. 2011. *The Kaldi Speech Recognition Toolkit*. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society. Available online at http://publications.idiap.ch/downloads/papers/2012/Povey_ASRU2011_2011.pdf

Robinson, T. 1997. *BEEP – British English example pronunciations*. Available online at http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html

Weide, R.L. 1998. *The Carnegie Mellon pronouncing dictionary*. Available online at http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Young, S. et al. 2009. *The HTK Book (for HTK Version 3.4)*. N.p.: Microsoft Corporation/Cambridge University Engineering Department.

Linguistic Development of the Alberta Bituminous Sands

Caelan Marrville
University of Alberta
caelan@ualberta.ca

Antti Arppe
University of Alberta
arppeualberta.ca

In public discourse concerning the Alberta bitumen sands located in northern Canada, there are two popular terms used to refer to the region: the *oilsands* and *tarsands*. Although these terms refer to the same area of North-Eastern Alberta, they currently have two very different semantic frames, the latter used largely by pro-environmental groups, and the former used presently by the Canadian federal and Albertan provincial governments as well as international oil companies. While the general connotations of these two opposing terms are known, how they are actually realized in texts, and how they evolved over time, has not been researched.

Given the number of controversial environmental and geopolitical issues related to the development of the bitumen sands industry, both proponents and opposition alike desire the most appropriate frame to express their message. As explained by Cosh (2012), not only does opting to use *oilsands* avoid the ugly associations of *tar*, but it also keeps the focus on *oil*, which is internationally associated with commerce. The connotations associated to the *oilsands* and *tarsands* frame the bitumen sands industry in a particular light, and the preference to utilize one term over the other impacts the frame in which the hearer interprets the discourse.

The overall goal of this study is to outline a systematic operationalization of the theoretical concept of framing by applying, adapting and developing further quantitative research methods and data collections traditionally used in corpus and computational linguistics, and in particular concerning phenomena that have a clear cross-disciplinary interest in the social sciences. The practical goal of our work is the study of alternative frames on controversial topics in public discourse. This will be demonstrated with data collected from a corpus created from provincial debate records dating back to the 1970's to trace the historical evolution of the conceptual associations of the two bitumen sands terms that are competing for the representation of economic, social, and environmental activity taking place in the Alberta oil industry.

We approach frame semantics through the model of Valence Framing Effects (Levin and Gaeth 1988; Levin et al. 1998). Levin et al. explain that a frame is capable of casting the same information about a phenomenon in either a positive or negative light and that a frame has both cognitive and motivational consequences in determining our perception. The valence of a frame and how it casts a particular phenomenon positively or negatively can therefore provide a basic but crucial perspective towards the applied frame. In an analysis of valance framing effects, determining the negative or positive association a frame has is essential to understanding both the function of a frame and how it may be interpreted.

Previous work on semantic framing has largely focused on the manual *ad hoc* analysis of a small set of written or spoken texts motivated by one side of an argument or another. Recent studies in framing and conceptual metaphor have taken new approaches (Heylen et al. 2008, 2013; Jaspaert et al. 2011; Peirsman et al. 2010; Magnusson et al. 2005). Nonetheless, there is an absence of purely empirical studies on semantic framing. It is our goal to bridge a quantitative approach to the study of cognitive semantics. Using a quantitative linguistic approach – namely collocation – we identify patterns of semantic framing that go beyond the scope of individual observations. Using collocational networks, we identify the psychological valence of collocates of each of the bitumen sand frames.

Collocational networks are created of the most significantly mutually co-occurring words between specified targets and the corpus. These networks are, in theory, collectively representative of the most significant themes surrounding a particular topic. In our case, the selection of *oilsands* and *tarsands* as the target nodes allow us to identify strongly associated terms within the provincial Hansard records. The collocations can characterize the most common strategies employed in the communication of information regarding the bitumen sands industry within the provincial governments. It follows that an analysis of these collocational networks could be used to understand the most relevant terms and expressions used in discourse on the bitumen sands within their respective provincial governments.

Collocational networks were created using pointwise mutual information scores and log likelihood ratios. Looking at the collocational networks through a visual categorical analysis, it was found that within Alberta, the *oilsands* and *tarsands* terms have similar collocates within the categories of current and future development, the raw resources, and production. The *oilsands* collocates within British Columbia shared more similarities with the Alberta data and shared collocates within economics, future development, contamination and raw resources. Notably the *tarsands* British Columbia shared little with either the Alberta collocates or British Columbia *oilsands* data.

A second analysis was carried out using norms of valence, a dataset containing normative emotional ratings for English words based on three dimensions: affective valence, arousal and dominance (Warriner et al. 2013). Given our interest in the positive and negative associations for our collocational networks, our analysis of the norms of valence data focused on the affective valence scores, which identify the level of positivity or negativity a given word carries. Linear model estimation using ordinary least squares (within the rms package in R) was calculated to test between the different variables. The calculations showed that *Term* (*oilsands* vs *tarsands*) and *Decade* had the greatest effect on the affective valence (pr(>t) = 0.0001), while *Province* only had a significant effect for the Alberta data.

We fit the data to a linear regression model to look at affective valences of collocates for each province individually. Within the Alberta subset, the mean affective valence score fell for collocates, increasing in negative valence over each decade for both the *tarsands* and *oilsands* data (Pr(>\t\) <0.0001 and Pr(>\t\) <0.0005, respectively). In the British Columbia data, *tarsands* also decreased in affective valence over time (Pr(>\t\) <0.0011), while *oilsands* collocates increased in positive association, albeit not to a degree of statistical significance (Pr(>\t\) <0.645). Within the British Columbia subset the affective valence decreased over time, gaining a slightly more negative valence within the *tarsands* data, while the *oilsands* data showed a positive increase in association. Figure 1.7 and 1.8 visualize the results of the linear regression models for the subsets of data.

Within Alberta, the mean valence was more negative for *tarsands* than for *oilsands*, and while the overall behavior of the two terms was found to be relatively similar, both *tarsands* and *oilsands* showed decreases in valence over time. This lead to the conclusion that, at least in the province of Alberta, the two terms are used in similar contexts. On the other hand, there did appear to be a difference between the *oilsands* and *tarsands* terms within the British Columbia corpus materials. The linear regression showed that diachronically the mean valence of *oilsands* collocates became more positive, while the valence of the *tarsands* collocates became more negative. In the British Columbia data, *tarsands* seem to indeed carry the stereotypical negative association, while *oilsands* is used in more positive contexts.

These results corroborate the findings of a visual data inspection from our previous analysis. The British Columbia *tarsands* collocational network differs from the British Columbia *oilsands* and Alberta *tarsands* and *oilsands* networks. In the British Columbia *tarsands* network, negative associations, references to *'dirty oil'* and concern over environment and contamination are common and therefore cause a lower affective valence for the term.

Attribute framing is considered an elementary approach to semantic framing where one characteristic of a phenomenon serves as the primary focus of the frame manipulation. As Levin & Gaeth (1988) suggest, attribute framing occurs because information is encoded relative to its descriptive valence – that positive labeling of an attribute leads to an encoding of information that evokes favorable associations, whereas the negative labeling of an attribute is likely to cause an encoding that evokes unfavorable associations in cognitive processing. The two alternative terms for the bitumen sands: *tarsands* and *oilsands,* serve as an excellent example of this kind of frame manipulation. On the one hand you have *oilsands* – a term which highlights the end goal and purpose of the extraction process and has connotations of industriousness and capital gain, on the other you have *tarsands* – a term that highlights the physical appearance of the extracted product prior to industrial processing and carries connotations of dirtiness and pollution. The two terms reflect different aspects of the same object and from an attribute framing perspective, the manipulation of these objectively equivalent terms is designed to accentuate their positive and negative connotations.

References

Cosh, C. (April 3, 2012). Don't call them 'tarsands'. *Maclean's.* Retrieved from http://www.macleans.ca/news/canada/oil-by-any-other-name/

Heylen, Kris, Thomas Wielfaert, and Dirk Speelman (2013). *Tracking Immigration Discourse through Time: A Semantic Vector Space Approach to Discourse Analysis*.

Jaspaert, Koen, et al. (2011). Does framing work? An empirical study of Simplifying Models for sustainable food production. *Cognitive Linguistics* 22.3, 459-490.

Levin, I. P., & Gaeth, G. J. (1988). How consumers are affected by the framing of attribute information before and after consuming the product. *Journal of consumer research*, 374-378.

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational behavior and human decision processes*, *76*, 149-188.

Magnusson, C., Arppe, A., Eklund, T., Back, B., Vanharanta, H., & Visa, A. (2005). The language of quarterly reports as an indicator of change in the company's financial status. *Information & Management*, 42, 561-574.

Peirsman, Y., Heylen, K., & Geeraerts, D. (2010).

Applying word space models to sociolinguistics. Religion names before and after 9/11. *Advances in Cognitive Sociolinguistics, Cognitive Linguistics Research [CLR]*, 111-137.

Warriner, A.B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior Research Methods, 45, 1191-1207.

Quite + ADJ seen through its translation equivalents: A contrastive corpus-based study

Michaela Martinková
Palacký University
michaela.martinkova@upol.cz

1 Introduction

Quite is a polyfunctional modifier, i.e. it belongs to two opposing intensifier types, in Biber et al. (2007, 556) called amplifiers/intensifiers and diminishers/downtoners, in Quirk et al. (1985, 590, 598) amplifiers (maximizers and boosters) and downtoners (compromizers), and in Paradis (1997, 27-28) reinforcers (maximizers and boosters) and attenuators (moderators).

Corpus linguists have dealt with *quite* in various syntactic positions both from a diachronic (Ghesquière 2012) and synchronic perspective (for the use of multivariate statistics see Desagulier 2012), and predominantly within the cognitive linguistics framework (e.g. Paradis 2008, Palacios 2009, Diehl 2003). While the studies mentioned above focus predominantly on British English, Levshina (2014) carries out a cluster analysis of the function of *quite* in the pre-adjectival position in twenty varieties of English, followed by a distinctive collexeme analysis of the main clusters; since the phonetic realization of *quite* cannot be taken into consideration, it is assumed that the function of *quite* directly follows from the semantic class of the modified adjective (combinations with individual degree modifier types are used as a diagnostics): *quite* modifying a limit adjective (*quite different*) or an extreme adjective (*quite brilliant*) is identified as a maximizer, in the premodification of a scalar adjective it is identified as a booster or a moderator (*quite good*).[78]

To avoid reliance only on the semantics of the collocating item, this study will adopt a different methodology: to identify the functions of *quite* modifying a predicative adjective, I will turn to a parallel translation corpus (Intercorp, version 7), i.e. I will systematically exploit "the bilingual intuition of translators, as it is reflected in the pairing of source and target language expressions" (Johansson 2007, 52). The varieties to be investigated is British (BrE) and American English (AmE) and the

[78] Since it is hard "to find formal contextual cues that can help one pinpoint the subtle differences between quite as a moderator "fairly, rather" and as a booster "very"" (both modify a scalar adjective), Levshina investigates only contexts in which *quite* is contrasted with the booster *very.*

language through which the meanings of *quite* is seen at this stage is a typologically different language, namely Czech.

2 Data and methods

Quite in the pre-adjectival position was retrieved from two subcorpora created within Intercorp, one of BrE (2,229,582 text positions) and one of AmE (3,170,937 text positions). 409 tokens of *quite* modifying a predicative adjective were identified in BrE and 158 in AmE, which means a statistically significant difference between the two varieties (LL 7932.66), namely the fact that *quite* in this position is much more frequent in BrE (183.4 pmw) than in AmE (49.8 pmw). Correspondences in the Czech translations were identified and sorted; for Czech degree modifiers, Ocelák's (2013) semantic classification (based on Paradis 2008) was used. Since "the same organization [of degree modifiers] applies in a similar way in both environments", i.e. in modification of adjectives and verbs (adverb-subjunct) (Paradis 1997, 23-24), all cases in which (due to typological differences) the pre-adjectival *quite* following a copula is rendered in Czech as a modifier of the verb or adverb were included.

3 Discussion of findings

In both varieties of English *sure* is the most frequent adjective (8.8% of all the tokens in BrE and 11.4% in AmE), followed in BrE by *different, right, clear, impossible, certain* and in AmE by *certain,* which seems to be in agreement with Levshina's observation that in L1 varieties, *quite* is typically used with "epistemic adjectives".

Unlike her research, however, the data show a higher proportion of a maximizer reading (translations with a maximizer or a minimizer, i.e. negative maximizer[79]) in BrE (32.3%) than in AmE (20%); such a big difference can hardly be attributed only to the fact that *quite* tends not to be used with extreme adjectives in AmE.[80]

In both varieties, the ratio of tokens with a reinforcing reading is further increased by tokens translated by expressions classified as emphasizers (BrE 4.1%, AmE 3.8%), both by those commenting upon the appropriate use of the expression (*quite alarming - doslova alarmující* [literally alarming]) and those expressing a comment "that what is being said is true" (Quirk et al. 1985, 583), cf. e.g. *quite good - opravdu dobrý* [really good]. Boosters (e.g. *velmi/velice* [very], *pěkně* [nicely], *děsně* [horribly]), cover 10.1% of the data in AmE, but they are also found in the translations of British fiction (4%). Boosters in correspondences of *quite* modifying an extreme or limit adjective seem to suggest that though a predicate with a maximizer modifying these adjectives might be semantically equivalent to a predicate with a non-modified adjective (cf. Ocelák 2013, 124), the "maximizer" paradoxically first attenuates the meaning of the adjective by imposing a scale into it and then it reinforces; an adjective with a lower intensity is sometimes found in the translation: (*quite striking – až omamně krásná* [even intoxicatingly beautiful]). The fact that the difference between a limit/extreme adjective modified by a maximizer and a non-modified adjective of this type is pragmatic rather than semantic (Ocelák ibid) is further reflected in the use of the Czech particles *teprve*, *jen* (both add expressivity to the sentence), hedging devices (cf. ex. 1), and also a high proportion (higher in BrE (31.1%) than in AmE (24.1%)) of zero correspondences (cf. ex. 2):[81]

(1) *It sounds quite rude*
Zní to nějak sprostě.
[sound:3SG it:NOM somehow rude:ADV]

(2) *Knut was quite right.*
Knut měl pravdu.
[Knut had: PTCP.SG.M truth:ACC]

The fact that a moderator (*celkem, poměrně, relativně*) is only found in 5.7% tokens of BrE and 4.4% tokens found in AmE does not have much informative value since additional 8.8% (BrE) and 12.7% (AmE) of the tokens were translated by the modifier *docela*, which is ambiguous between a maximizer and a moderator reading, and 5.9% (BrE) / 9.5% (AmE) by *dost,* which is another ambiguous degree modifier (moderator/booster). In addition, three tokens (BrE 2 and AmE 1) of an approximator (*téměř*, *takřka* [almost]) and five (BrE 1 and AmE 4) of a diminisher (*poněkud* [somewhat]) were found.

4 Conclusions and future prospects

My parallel corpus-based study confirms a pragmatic rather than semantic function of *quite* in the premodification of predicative adjectives (a high percentage of zero correspondences, translations with hedging devices and particles) and reveals a rich polysemy network synchronically observed. As

[79] See Quirk et al. (1985, 597). My example is *It was quite irrelevant. Není vůbec podstatné* [be:NEG.3SG at-all relevant].

[80] It follows from Levshina's charts that in BrE limit adjectives cover about 25% of the tokens of *quite*+ADJ and extreme adjectives about 4.2% (in AmE 32% and 2.9% respectively).

[81] Aijmer makes a similar argument about English discourse particles: "discourse particles in English do not affect the truth conditions of the utterance and do not add anything to its propositional content. Omission is therefore a possible translator strategy" (Aijmer and Altenberg 2002, 33).

to the difference between the British and American data, I could not confirm a higher percentage of the maximizer reading in AmE. To confirm a higher frequency of the moderator reading in BrE than in AmE (cf. Levshina 2014), more research has to be done on the role of ambiguous Czech degree modifiers *docela* and *dost,* which cover a significant percentage of (especially American) data.

References

Aijmer, K. and Altenberg, B. 2002. "Zero translations and cross-linguistic equivalence: evidence from the English-Swedish Parallel Corpus." In: L. E. Breivik and A. Hasselgren (eds), *From the COLT's mouth ... and others. Language corpora studies in honour of Anna-Brita Stenström.* Amsterdam: Rodopi.

Biber, D. et al. 2007. *Longman Grammar of Spoken and Written English*. Pearson Education ESL.

Desagulier, G. "Quite new methods for a rather old issue: Visualizing constructional idiosyncrasies with multivariate statistics." Available at http://www2.univ-paris8.fr/desagulier/home/handout_ICAME_33.pdf

Diehl, H. 2005. "Quite as a degree modifier of verbs." *Nordic Journal of English Studies* 4, (1), 11–34.

Ghesquière, L. 2012. "On the development of noun-intensifying quite." Paper presented at ICAME 33 conference.

Johansson, S. 2007. "Seeing through multilingual corpora". In R. Facchinetti (ed.) *Corpus Linguistics 25 Years On*. Amsterdam – New York: Rodopi.

Levshina, N. 2014. "Geographic variation of quite + ADJ in twenty national varieties of English: A pilot study." In: A. Stefanowitsch (ed.), *Yearbook of the German Cognitive Linguistics Association* 2 (1), 109-126.

Ocelák, R. 2013. "Sémantické škály a skalární modifikátory v češtině." *Slovo a slovesnost* 74 (2), 110-134.

Palacios Martínez, I. M. 2009. "Quite Frankly, I'm Not Quite Sure That it is Quite the Right Colour. A Corpus-Based Study of the Syntax and Semantics of Quite in Present-Day English. *English Studies* 90 (2), 180-213.

Paradis, C. 1997. *Degree modifiers of adjectives in spoken British English* (Lund Studies in English 92). Lund: Lund University Press.

Paradis, C. 2008. "Configurations, construals and change: expressions of DEGREE." *English Language and Linguistics* 12, 317-343

Quirk et al. 1985. *Comprehensive Grammar of the English Language*. London: Longman.

Czech National Corpus - InterCorp. Institute of the Czech National Corpus FF UK, Praha./

The Czech "modal particle" prý: Evidence from a translation corpus

Michaela Martinková
Palacký University
michaela.martinkova@upol.cz

Markéta Janebová
Palacký University
marketa.janebova@upol.cz

1 Introduction

In the last twenty years, contrastive linguistics has benefited greatly from the introduction of translation corpora; they supply the valuable "bilingual output" which "provides a basis of comparison, or at least justifies the assumption of comparability" (Gast 2012). This paper adopts the methodology of "seeing through multilingual corpora" (Johansson 2007) to identify the functions of the Czech "modal particle" *prý* in the genres represented in the multilingual translation corpus InterCorp.

Historically, *prý* is a reduced form of the reporting verb *pravit* [to say], namely its 3rd person sg. or aorist form (*praví* [say:PRS.3SG] or *pravi* [say:AORIST]) (Machek 2010, 481). Fronek (2000) differentiates two functions of the present-day *prý*: first, to introduce what other people say (in which case it is equivalent to *allegedly*, *reportedly*, *supposedly*, *apparently*, *they say*); second, to cast doubt on it:

(1) *Prý* [PART] o [about] tom [it] nic [nothing] neví [know:NEG.PRS.3SG].
He claims he does not know anything about it.

Czech monolingual dictionaries list one more function, namely the introduction of direct speech, arguably the only use in which *prý* does not carry modal overtones; *prý* is defined as a modal particle with the meaning of uncertainty and doubt caused by the fact that the information is only second-hand (Mejstřík et al. 2009).

Unlike the dictionary makers, we believe that considering the expression of speaker's doubt about the truth of what is being said as an inherent part of the meaning of *prý* (reference rather than inference) is, without systematic linguistic research, rather premature. According to Grepl (2002, 375), for example, *prý* is a third type of "reproducing original utterances", alongside direct and indirect speech, i.e., no doubt has to be present. In their study of the collocational profile of *prý* in the monolingual written SYN2000 corpus, Hoffmanová and Kolářová mention the important role *prý* has in the rendering of dialogues in fiction (2007, 98), recognizing the fact that the particle is more frequent in journalistic

texts than in fiction (101). On the other hand, Hirschová and Schneiderová's study of the adverb *údajně* [reportedly], which focuses on journalistic texts, stresses the notion of distance from the reported facts, motivated by an effort to avoid responsibility for the truth of reported statement, or to show disagreement with it. Both *prý* and *údajně* are considered as "lexical evidential markers", more specifically markers of reported evidentiality of the hearsay type, where the source of the reported information is not known (2012, 2).

The present study takes both papers as a starting point: we investigate the presence/absence of the source of the reported information in the English correspondences of *prý* in the genres represented in InterCorp (version 6).

2 Data

The genres represented are fiction, journalistic texts (the Presseurope database), and an approximation to the debate in the European Parliament (as documented in the proceedings – Europarl). The subcorpora created are not comparable, neither in size, nor the source language: while in the subcorpus of fiction texts Czech or English is always the language of the original, the same cannot be said about the Presseurope subcorpus, where the information about the language of the original text is missing in the metadata. Europarl questions the very concept of the source language: while "until 2003 the texts were translated directly from the source languages into any of the target languages ... [f]rom 2003 onwards … all languages were first translated into English and then into the relevant target language" (Gast and Levshina 2014). To obtain a sensible amount of Europarl data, a decision was made to include even translations from other languages.[82] Table 1 presents the sizes of the individual subcorpora (in text positions – TPs) and absolute/relative frequencies of *prý* in each subcorpus:

	Czech target texts (TTs)		Czech source texts (STs)	
	size	*prý* f/ipm.	size	*prý* f/ipm.
fiction	3,548,005	192/54.11	678,818	101/177.63
Press-europe	281,461	6/21.32	59,111	20/338.35
Euro-parl	size: 15,038,876 f/ipm.: 51/3.39			

Table 1: The size of the subcorpora created and the absolute (f)/relative (ipm) frequencies of *prý*

3 Discussion of findings

Table 2 suggests a difference between fiction, where the source of the reported information tends to be expressed, and Europarl, in which it is left unexpressed in the majority of tokens; Presseurope is in-between.

	Cz TTs		Cz STs	
	source known	source unknown	source known	source unknown
fiction	64.6%	35.4%	61.4%	38.6%
Presseurope			45%	55%
Europarl	source known: 17.6% source unknown: 82.4%			

Table 2: The percentage of the known/unknown source of reported information in the correspondences of *prý*.

A closer look at the fiction data reveals that the most frequent translation equivalent of *prý* is a reporting or parenthetic clause with the verb *say*. The subject of the reporting clause tends to be the 3rd person singular pronoun *he* or *she*, or a noun phrase with a definite reference. *Prý* is very often used to divide long reported segments. Overall, the data seem to suggest that in the texts of fiction *prý* predominantly has a reporting function.

In the Europarl texts, on the other hand, the most frequent translation equivalent of *prý* (51%) is a reporting clause with the verb in the passive voice, i.e., the source of the reported information is demoted. Adverbs, most frequently *allegedly*, *apparently*, *reportedly*, and *supposedly*, cover an additional 25.5% of the correspondences. The source of the reported information is explicitly mentioned in 10 tokens; unlike in the text of fiction, however, it is either too general ("*media reports*", "*people around me*"), or it is mocked.

In the Presseurope texts correspondences without the source of the reported information outnumber those with the source specified. However, the difference is smaller than expected; we have not proved that *prý* is used by journalists to avoid responsibility for what they are reporting (cf. Hirschová and Schneiderová 2012). Arguably, this can be explained by the specific nature of the texts included in Presseurope: they do not report news but tend to be polemic; reference is often made to authors of other articles (cf., e.g., "*at least, so says Karel Kříž*"). In the remaining cases of a known source, a wider context proves the reported statement to be either false, or at least open to discussion.

4 Looking Ahead

We believe that the translation equivalents of *prý* point to different primary functions of *prý* in the genres analyzed; while in the texts of fiction *prý* tends to report, in the Presseurope texts the authors either make reference to another text, or, like in the

[82] This strategy was also adopted in Gast and Levshina (2014).

Europarl texts, they express doubt about the reliability of the reported information.

This does not mean, however, that *prý* is never used to cast doubt on the reported information in the texts of fiction; sometimes the verbs *claim* is used in the reporting clause equivalent to *prý*, or a wider context makes it explicit that the reported information is false. In most cases, the sentence equivalent to the one introduced by *prý* is backshifted:

(3) (My mom never went to the circus) because *she said* it was too hot. (FICTION Day)
protože [because] je [it is] tam [there] *prý* [she-said] moc [much] horko [hot].

However, the presence of backshift can hardly be taken for a signal of the fact that the reported information is false. The mechanism which triggers such an interpretation, we believe, is pragmatic inferencing in the sense of Hopper a Traugott (2003, 79): "Grammaticalization changes seem to draw primarily on the second maxim of Quantity, in the form 'Say no more than you must and mean more thereby' [...] and Relevance"; if the speaker in (3) identified herself with her mother's statement, the reporting clause (or *prý*) could well be left out. Our analysis tentatively suggests that the inference is obligatory if the source of the reported information is a direct participant in communication (namely the listener); however, a subtle description of the mechanism is left for future research.

References

Czech National Corpus - InterCorp, Institute of the Czech National Corpus, Prague. <http://www.korpus.cz>.

Fronek, J. 2000. *Velký česko-anglický slovník*. LEDA.

Gast, V. 2012. "Contrastive Linguistics: Theories and Methods." <http://www.personal.uni-jena.de/~mu65qev/papdf/contr_ling_meth.pdf>

Gast, V. and Levshina, N. 2014. "Motivating w(h)-clefts in English and German: A hypothesis-driven parallel corpus study." <http://www.personal.uni-jena.de/~mu65qev/papdf/gast_levshina_subm.pdf>

Grepl, M. 2002. Reprodukce prvotních výpovědí. In: P. Karlík et al. (eds.), *Encyklopedický slovník češtiny*. Prague: Nakladatelství Lidové noviny.

Hirschová, M. and Schneiderová, S. 2012. Evidenciální výrazy v českých publicistických textech (případ údajně–údajný). [online]. <http://www.ujc.cas.cz/miranda2/export/sitesavcr/data.avcr.cz/humansci/ujc/vyzkum/gramatika-a-korpus/proceedings-2012/konferencni-prispevky/HirschovaMilada_SchneiderovaSona.pdf>

Hoffmanová, J. and Kolářová I. 2007. "Slovo *prý/prej*: možnosti jeho funkční a sémantické diferenciace." In: F. Štícha and J. Šimandl (eds.), *Gramatika a korpus 2005*. Praha: ÚJČ AV.

Hopper, P. and Traugott, E. C. 2003. *Grammaticalization*. Cambridge: CUP.

Johansson, S. 2007. *Seeing through Multilingual Corpora. On the Use of Corpora in Contrastive Studies*. Amsterdam/Philadelphia: John Benjamins.

Machek, V. 2010. *Etymologický slovník jazyka českého*. Prague: Nakladatelství Lidové Noviny.

Mejstřík, V. et al. (eds). 2009. *Slovník spisovné češtiny pro školu a veřejnost*. Prague: Academia

Semantic Word Sketches

Diana McCarthy
Theoretical and Applied Linguistics
Univ. Cambridge
diana@dianamccarthy.co.uk

Adam Kilgarriff
Lexical Computing Ltd.
adam.kilgarriff@sketchengine.co.uk

Miloš Jakubíček
Lexical Computing Ltd.
Masaryk University
milos.jakubicek@sketchengine.co.uk

Siva Reddy
Univ. Edinburgh
siva.reddy@ed.ac.uk

A central task of linguistic description is to identify the semantic and syntactic profiles of the words of a language: what arguments (if any) does a word (most often, a verb) take, what syntactic roles do they fill, and what kinds of arguments are they from a semantic point of view: what, in other terminologies, are their selectional restrictions or semantic preferences. Lexicographers have long done this 'by hand'; since the advent of corpus methods in computational linguistics it has been an ambition of computational linguists to do it automatically, in a corpus-driven way, see for example (Briscoe et al 1991; Resnik 1993; McCarthy and Carroll 2003; Erk 2007).

In this work we start from word sketches (Kilgarriff et al 2004), which are corpus-based accounts of a word's grammatical and collocational behaviour. We combine the techniques we use to create these word sketches with a 315-million-word subset of the UKWaC corpus which has been automatically processed by SuperSense Tagger (SST)[83] (Ciaramita and Altun 2006) to annotate all content words with not only their part-of-speech and lemma, but also their WordNet (Fellbaum 1998) lexicographer class. WordNet lexicographer classes are a set of 41 broad semantic classes that are used for organizing the lexicographers work. These semantic categories group together the WordNet senses (synsets) and have therefore been dubbed `supersenses' (Ciaramita and Johnson, 2003). There are 26 such supersenses for nouns and 15 for verbs. Table 1 provides a few examples and the full set can be seen in Ciaramita and Altun, (2006). SST performs coarse word sense disambiguation, to identify which WordNet supersense a word belongs to. We note that this is not a case where WSD accuracy is critical. In the spirit of Kilgarriff (1997), it is 'background' WSD used for developing a lexical resource. It is hoped that, to a large extent, individual errors in disambiguation are filtered out as noise by the signal from the correct cases.

Noun Supersense	**Nouns denoting**
act	acts or actions
animal	animals
artifact	man-made objects
...	
Verb Supersense	**Verbs of**
body	grooming, dressing and bodily care
consumption	eating and drinking
communication	telling, asking, ordering, singing
...	….

Table 1: Some example supersense labels and short descriptions from the WordNet documentation

Each entry in a word sketch shows a different combination of supersenses within a syntactic relationship. As an example, Figure 1 shows a semantic word sketch for the English verb *fly*. The semantic word sketches are produced by a bespoke 'sketch grammar' which identifies the grammatical and collocational behaviour using the part-of-speech tags and finds the predominant semantic classes of the arguments in the syntactic relationships using the supersenses associated with the words in the corpus.

The sketch is presented in tables where the column header states the syntactic pattern identified (e.g. *intransframe*). Then, within each pattern, the head arguments are indicated by the supersense labels (with a part-of-speech suffix) with an asterisk indicating the supersense of the target word in each case (***motion** in the examples for *fly* in Fig. 1).

The first and most salient table, *intransframe* lists intransitive frames with the first being **animal.n_*motion.v** – where the verb has an animal subject. There were 392 hits, with a logdice[84] salience score of 10.12. Clicking on the number, we can explore the concordances, as shown in Figure 2. As can be seen, these are valid instances of this frame.

83 SST is available at http://sourceforge.net/projects/supersensetag/

84 See Rychlý 2008.

fly *(verb)*

UKWaC super sensed freq = 22,610 (61.1 per million)

intransframe	4,536	**8.5**
animal.n_*motion.v	392	10.12
artifact.n_*motion.v	1,007	9.58
time.n_*motion.v	240	8.8
person.n_*motion.v	1,323	8.36
communication.n_*motion.v	213	8.2
group.n_*motion.v	285	7.65
act.n_*motion.v	166	7.63
0_*motion.v	100	7.56

mwe	1,750	**0.6**
fly_by_motion.v	413	12.33
fly_on_motion.v	291	11.99
fly_start_motion.v	194	11.54
fly_colours_act.n	141	11.15

transframe	1,074	**4.1**
person.n_*motion.v_artifact.n	103	8.81

ne_subject_of	974	**2.1**
*_motion.v	892	8.31

caternative	551	**2.2**
*motion.v_motion.v	177	8.31

Figure 1: Semantic word sketch for English *fly*.

Next comes the **artifact.n_*motion.v** i.e. artifact-as-subject frame, top lemmas in the subject slot being *plane, flag, ball, aircraft, helicopter, shot, airline, bullet.* It is not immediately apparent if this is a separate sense of *fly* to animal-as-subject: it depends on whether the user (such as a lexicographer) making the choice is a 'lumper' or a 'splitter': how fine-grained they like their senses to be.[85] It is also not clear whether *plane* (which is self-moving) fits in the same sense as *ball* (which is not; or *flag,* possibly an intermediate case).

85They may also be working to other constraints, like never giving the same sense where a key argument has a different lexicographer class.

graphics like ***birds***	**flying**	(inspired from the
spectacle of a Harris ***hawk***	**flying**	around the building
been showing ***birds***	**flying**	on and off of a wire
with all the ***birds***	**flying**	away , which fits into
appeared , showing ***birds***	**flying**	and rivers trickling
climaxed , all the ***birds***	**flew**	off in unison. it was
before the ***butterfly*** can	**fly**	. Adult The main goal
, like the ***vulture***	**flying**	on high , he saw the
, like the ***vulture***	**flying**	on high , he saw the
pond . A wood ***pigeon***	**flies**	up to the oak with
background) . The ***female***	**flew**	up onto the cotoneaster
upwards to three ***cranes***	**flying**	in a V-formation from
The ***nuthatch*** is still	**flying**	in to feed on the sunflower
sand . A ***cormorant***	**flew**	along offshore while
Tortoiseshell ***butterfly***	**flies**	up the garden and over
A green ***woodpecker***	**flies**	down to the track ahead

Figure 2: concordance (truncated to fit) for **animal.n** as subject of *fly.v.*

time.n_*motion.v relates to the idiom of time flying past, with lemmas in addition to *time* itself being *day, night, hour, year* and *winter.*

The next table, **mwe** (multiword expresssions), covers two prepositional verbs and two idioms: *a flying start* and *flying colours,* which cannot usefully (from a semantic, or lexicographic, point of view) be treated elsewhere in the entry.

The third table, **person.n_*motion.v_artifact.n**, for transitives, here has just one frame (over the frequency threshold; here set at 100). These are mostly people flying aircraft, with occasional instances, missed by the *mwe* filter, of people flying the flag.

The fourth table and its single frame cover cases where named entities (*ne*'s – named people and organisations) are doing the flying. The fifth and final one covers some problematic parsing cases: *fly tipping* (a British idiom meaning depositing rubbish illegally), *fly casting* and *fly reel* (technical terms from the fishing domain) and *flying trapeze* (from the circus).

5 Formalism

The formalism in which the sketch grammar (which specifies the word sketch) is written is an extended and augmented version of the one described in Kilgarriff et al. (2004), itself adopted and adapted from Christ and Schulze (1994). The full documentation is available at the Sketch Engine website.[86]

86 http://www.sketchengine.co.uk

6 Relation to FrameNet and similar projects and to distributional semantics

FrameNet (Baker et al. 1998) has been probably the most ambitious and inspiring project for building a lexical resource in recent years. It aims to establish the set of semantic frames for the predicates of English, complete with a description of the semantic roles of the arguments (the frame elements) in each frame and their syntactic and semantic specifications. It has been a highly influential project, spawning a wide variety of subsequent projects, for example to produce FrameNets for other languages, to automate the process of building FrameNet-like resources, and to disambiguate words according to frames.

FrameNet is a manual project: people decide what the frames are, what instances fit which frames, and so forth, albeit with sophisticated computational support. This assures high accuracy, but also slow progress. The contribution that semantic word sketches might make to FrameNet-style projects (including Corpus Pattern Analysis, (CPA: Hanks 2013) and VerbNet (Kipper et al 2006)) could be

- helping in extending the coverage of the dataset by providing data to explore, edit and include
- providing an additional type of data that is not currently available within FrameNet: the actual argument slot fillers for a given frame with frequency data and corpus examples.
- research into what parts of, and to what extent, the FrameNet-style lexicography process can be automated, where manual entries provide a gold standard which semantic word sketches aspire to. One important aspect not covered by semantic word sketches is the role of the semantic arguments within a frame (for example whether the subject of *fly* is riding a vehicle or a self-mover).

Semantic word sketches offer the benefits that the corpora, WordNets (or their younger relation, Babelnet (Navigli and Ponzetto 2012)) and the computational tools needed to create them are already in place for a number of languages, so the investment needed to create them for a new language is modest.

Semantic word sketches contrast with FrameNet, VerbNet and CPA through being automatic. There is another stream of work with similar goals but that is completely data-driven and, unlike semantic word sketches, does not use a manually created resource (WordNet) for defining semantic classes. This stream is distributional semantics (see Baroni and Lenci (2010) for an overviewand Socher (2014) on the recent and related area of 'deep learning'). We look forward to exploring the contrasts and complementarities between semantic word sketches and distributional semantics.

References

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998, August). The Berkeley FrameNet Project. *Proc. ACL.*

M. Baroni and A. Lenci. 2010. Distributional Memory: A general framework for corpus-based semantics. Computational Linguistics 36(4): 673-721.

Ciaramita, M. and Altun, Y. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. Proc EMNLP, Sydney, Australia: pp 594-602.

Ciaramita M. and Johnson, M 2003. Supersense Tagging of Unknown Nouns in WordNet. In Proceedings of EMNLP 2003.

Erk K. 2007. A simple, similarity-based model for selectional preferences. Proc. ACL 2007. Prague, Czech Republic, 2007.

Fellbaum, C editor. 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge.

Hanks, P. W. 2013. Lexical Analysis: a theory of norms and exploitations. MIT Press.

Kilgarriff, A. 1997. Foreground and Background Lexicons and Word Sense Disambiguation for Information Extraction. Proc Workshop on Lexicon-driven Information Extraction, Frascati, Italy.

Kilgarriff, A., Rychlý, P., Smrz, P., Tugwell, D. 2004. The Sketch Engine. Proc. EURALEX. pp. 105–116.

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. 2006. Extending VerbNet with novel verb classes. *Proc. LREC*.

McCarthy, D. and Carroll J. 2003. Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences, *Computational Linguistics,* 29(4). pp 639-654.

Navigli. R., and S. Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, Elsevier, pp. 217-250.

Resnik. P. 1993. Selection and Information: A Class-Based Approach to Lexical Relationships. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Rychlý, P. 2008. A Lexicographer-Friendly Association Score. Proc. RASLAN workshop, Brno, Czech Republic.

Schulze, B. M., & Christ, O. 1994. The CQP user's manual. *Universität Stuttgart, Stuttgart*.

Socher, R. 2014. Recursive Deep Learning for Natural Language Processing and Computer Vision, PhD Thesis, Computer Science Department, Stanford University

Twitter rape threats and the Discourse of Online Misogyny (DOOM): using corpus-assisted community analysis (COCOA) to detect abusive online discourse communities

Mark McGlashan
Lancaster University
m.mcglashan@lancaster.ac.uk

Claire Hardaker
Lancaster University
c.hardaker@lancaster.ac.uk

1 Introduction

In July 2013, Caroline Criado-Perez successfully campaigned to have a woman appear on an English banknote, and was subsequently inundated with extreme misogynistic abuse on Twitter. Over the following days, this escalated, and more prominent women were sent death and bomb threats. Whilst legislative bodies came under intense pressure to handle the issue, there is a lack of research into such behaviour, making evidence-based solutions difficult. As a result of an ESRC urgency grant dedicated to investigating this case, this presentation outlines the project and some findings with regards to the methodological background to finding rape threat trolls and identifying the networks that they form, as well as some suggestions for future directions and projects.

2 Data and sampling

The work focuses on a database of interactions with the @CCriadoPerez Twitter account that occurred over a period of 3 months in 2013. The sample was restricted to only tweets which were sent to (i.e. mentioned) or from (i.e. tweeted from) the @CCriadoPerez account during the period 25 June 2013 to 25 September 2013.

The sample was collected using Datasift (www.datasift.com), a commercial service that provides access to 'social data' from sources such as Twitter and Tumblr. The 'historics' functionality allows users, unlike other services, to search for and easily collect millions of historical online posts.

Our sampling frame returned 76,235 interactions involving @CCriadoPerez: 67,129 mentions of the account and 9,106 tweets from the account. Converted into a corpus, the language data in these tweets (excluding URLs, hashtags and mentions) amounted to a total of 1,014,222 words, which we call the Caroline Criado-Perez Corpus (CCPC).

3 Methods: Corpus-assisted Community Analysis (COCOA)

Corpus-assisted Community Analysis is a multimethodological approach to the study of online discourse communities that combines methods from Discourse Analysis, Corpus Linguistics and Social Network Analysis.

3.1 Corpus-driven Discourse Analysis

Drawing predominantly on the work of Baker (2006), we took a corpus-driven approach to analysing discourses in the language contained in the CCPC. The CCPC was explored and analysed using the AntConc corpus analysis tool (Anthony 2014). By taking a corpus-driven approach the analysis of linguistic units like lexis and n-grams, we were able to gain an overall view of frequent topics and discourses that were present in conversations contextualized by a period of sustained online abuse focused on an individual online user.

3.2 Social Network Analysis

At the beginning of the research, we were informed by Caroline Criado-Perez, the target of the online abuse in question, about a number of threating and offensive online communications that she had received (and which initiated this research). In informing us, she disclosed the screen names of all of the accounts that she was aware of that were sending her abuse. As such, we began the research with a seed list of account names of abusive users.

Using these names, we were able to search the database of 76,235 tweets for all of the interactions in which those abusive users identified by Criado-Perez occurred (either as someone who tweets or is mentioned) and to populate a larger sample of abusive users which we categorise into two groups: high-risk users (n=61) and low-risk users (n=147). High-risk users exhibited behaviours such as, intent to menace, harassment, and sexual aggression. Low-risk users tended to make misogynistic or generally insulting remarks but were not considered threatening.

4 Corpus-assisted Community Analysis (COCOA): implementation and findings

Subsets of the database and CCPC were created based on finding from the implementation of methods discussed in 3.1 and 3.2 with regard to high-risk and low-risk users: a CCP high-risk corpus, CCP low-risk corpus, and CCP no-risk corpus. The CCP no-risk corpus is comprised of tweets from users not identified as being abusive.

4.1 Linguistic findings

Using keyword analysis, the language found in the CCP high-risk and CCP low-risk corpora was compared to that found in the CCP no-risk corpus which highlighted some of the key discourses in the talk of 'risky users', including:

- Rape: *rape, raep, raping*
- Misogyny: *cunt, bitch*
- Homophobia: *faggot, gay*
- Racism/anti-Semitism: *nigger, jew*
- Genitalia/anatomy: *pussy, penis, fucking, ass, cock*

Moreover, we found that the overall response by users of Twitter to the abuse received by Criado-Perez was to condemn abuse and to contest online abuse (especially misogynistic abuse).

4.2 Applying linguistic findings in detecting discourse communities

After having found what language was *key* about the CCP high-risk and CCP low-risk corpora, we were also interested in if 'risky' users associated, and how they associated if they did. We found that 'risky' users frequently associated with each other in directed networks (cf. Scott 2013) as well as through ambient affiliation (Zappavigna 2012). We assessed directed networks as existing when 'risky' users mention other 'risky' users in their tweets. Ambient affiliation (which can also be understood in terms of undirected networks (cf. Scott 2013)) was assessed as being instances in which 'risky' users talk about the same things regardless of whether they are known to each other or not.

We found that not only do 'risky' users mention each other (ergo, talking to one another and forming directed networks) but also collectively engage in the use of abusive language (relying on discourses outlined in 4.1) when targeting Criado-Perez.

5 Conclusions

We applied a combination of methods from Corpus Linguistics, Discourse Analysis and Social Network Analysis in order to find a community of users on Twitter who communally engage in the use of a range of offensive discourses in order to enact abuse online.

Corpus Linguistics enabled the detection of a number of discourses which were significant in the population of Twitter users that partook in a spate of online abuse.

Social Network Analysis enabled to us to populate a large sample of abusive users based on a smaller group of users identified as being abusive and to explore the networks and interactions in which 'risky' users were involved.

Our proposed combination of methods for Corpus-assisted Community Analysis (COCOA) gave us a framework for combining Corpus Linguistic and Social Network Analytical approaches to the analysis of online communities and brought together the findings of two different forms of analysis to provide a more complex and complete overview of the discursive and networking behaviour of 'risky' online users during a period of prolonged abuse in the context of Twitter.

Acknowledgements

This work was supported by the Economic and Social Research Council [grant number ES/L008874/1].

We would like to thank: Steve Wattam for his tireless help, engineering wizardry and endless cynicism; Uday Avadhanam for initiating him into the (horribly complex) world of R; Tony McEnery for all his help and patience.

References

Anthony, L. 2014. AntConc (Version 3.4.3.) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Scott, J. 2013. *Social Network Analysis*. 3rd Ed. London: Sage.

Zappavigna, M. 2012. *Discourse of Twitter and social media*. London: Continuum.

A Corpus Based Investigation of 'Techno-Optimism' in the U.S National Intelligence Council's Global Trends Reports

Jamie McKeown
Hong Kong Polytechnic University
Jamie.mckeown@gmail.com

1 National intelligence council

Established in 1979 the National Intelligence Council (NIC), a collection of leading academics and subject matter experts, convenes every four years in order to project how 'key trends might develop over the next decade and a half to influence world events' (NIC 2020). The findings of the NIC are compiled as a key policy document for each newly incumbent U.S. president. Since 1997 the 'non-sensitive' aspects of the NIC reports have been placed in the public domain.

Charged with the task of projecting 15 years ahead of the date of publication, consideration of technology is central to the work of the NIC as the ability to predict the future is an inseparable function of the ability to foresee technological development given that technology saturates the organizational base of any given society or era (Popper 1957).

2 The future and 'techno-optimism'

The global community currently faces unprecedented challenges in critical areas (e.g. food production, resource depletion, energy availability) with rhetoricians quick to portray technological innovation as the panacea to avoid systemic collapse. At times technology has arguably helped deliver the species from peril but the extent to which we can unquestioningly rely on technological innovation or the extent to which it actually increases risk of catastrophe is worryingly overlooked in popular discourse (Wright 2005; Taleb 2012).

Huesemann and Huesemann (2011) go so far as to boldly claim that humanity has seduced itself with ideological 'techno-optimistic' notions of salvation. They define techno-optimism in terms of a set of core beliefs:

- efficiency gains will solve the major problems of our day;
- continued economic growth is environmentally sustainable;
- military investment will ensure global peace and secure access to scarce resources for industrialised nations;
- high-tech solutions and medicinal drugs will abolish disease;
- biofuels and nuclear energy will replace fossil fuels;
- genetically modified (GM) crops will ensure food security;
- technology is an autonomous independent force that cannot be stopped nor censored.

As noted by Eagleton (1991) an ideology can 'be put into contradiction by imbuing it with a form which highlights its hidden limits, thrusts it up against its own boundaries and reveals its gaps and elisions' (1991: 46). Huesemann and Huesemann, in their exposition of techno-optimism, systematically draw upon evidence from various fields to expose the incoherencies within 'techno-optimism' as an ideology, such evidence includes:

- according to the laws of thermodynamics efficiency gains only work if demand for the limited resource is kept constant or outpaced by efficiency innovation (Jevons 1896);
- historically efficiency gains have often lead to an increase in consumption of limited resources through the price reduction of goods or an increase in disposable income (Greening et al. 2000);
- infinite growth is not possible on a finite planet (MacKellar 1996);
- nuclear arsenals will provide no defence against environmental catastrophe brought about by over-exploitation of resources (Commoner 1971) ;
- the greatest gains in medical science have been driven by two simple factors: nutrition and hygiene not medical breakthroughs based on reductionist science (McKeown 1979);
- whilst biofuels are associated with lower greenhouse emissions the aggregate environmental impact is actually greater than petrol fuels (Scharlemann and Laurance 2008);
- the need for genetic modification makes the presupposition that issues of food security are caused by inherent limits of the earth's capacity to feed its inhabitants as opposed to current pathological distribution systems (Runge et al. 2003; Pollon 2007);
- selection and prohibition of technology is, and always has been, conducted according to the interests of the politically dominant (Dickson 1975).

The primary purpose of this paper was to investigate the prevalence and operation of techno-optimism in the work of the NIC (2010-2035).

3 Data and approach

The five reports published thus far by the NIC were used as data to build a small corpus (140,000 words).

Wmatrix3 was used to extract all concordance lines that fell within the semantic domain of technology. The data was then analyzed from a number of perspectives. First, the concordance lines were thematically classified according to the propositional sub-categories of techno-optimism (outlined above) in order to gain an understanding of the focus of the NIC. The data was subsequently examined in terms of macro-propositional agreement (whether explicit or implicit) with the core set of values and beliefs. At the textual level, the concordance lines were examined for semantic prosody (Sinclair 1991) and finally, in order to gage the level to which the author's explicitly intruded upon the text in order to provide persuasive commentary the concordance lines were examined for the presence of epistemic and attitudinal stance markers (Hyland 2005a).

4 Preliminary results

1,043 concordance lines were extracted from the corpus; 94% of these were classified as relevant. Preliminary results revealed that *Military Investment* was the largest sub-category in the corpus (n = 34%). Perhaps the dominance of the *Military Investment* subcategory is a product of the military adventurism pursued by successive U.S governments and its allies since the turn of the century.

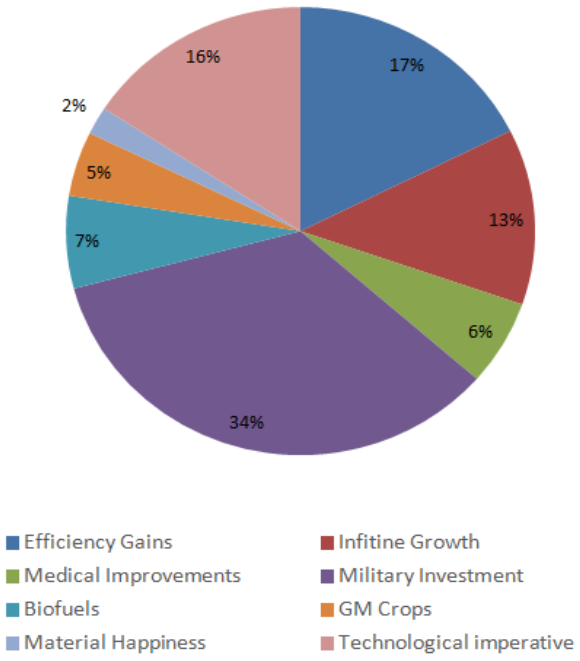


Figure 1: Distribution of Techno-Optimist Subcategories in the NIC Corpus.

Subcategory of Techno-Optimism	Agreement Level (%)
Efficiency Gains	0.92
Infinite Growth	0.97
Medical Improvements	0.97
Military Investment	0.78
Biofuels and Nuclear Power	0.84
GM Crops	0.93
Material Happiness	1.0
Technological Imperative	0.92
Overall Agreement level	0.89

Table 1: Techno-Optimist Macro-Propositional Agreement in the NIC Corpus

The high level of macro-propositional agreement suggests that techno-optimism was highly prevalent in the work of the NIC. In relation to *Efficiency Gains*, *Infinite Growth* and *Material Happiness* this disclosed a status quo bias (i.e. business as usual) in the work of the NIC. Radical possibilities such as the development of a seasonal economic system or localized economies were not given consideration.

Interestingly whilst *Military Investment* was the largest sub-category it also had the highest level of macro-propositional disagreement. This was largely driven by consideration of the challenges posed by new technologies to traditional power structures. In the remaining subcategories disagreement was mostly driven by consideration of technical and commercial constraints imposed upon the process of innovation diffusion.

Stance Marker	Items Per 1000 Words	% of Total Use
Uncertainty Marker (Hedge)	12.4	0.76
Certainty Marker (Booster)	1.2	0.08
Attitude Marker	2.5	0.16
Total	16.1	100

Table 2: Epistemic Stance Markers in the NIC Corpus (Semantic Field: Technology).

The reader may appreciate a benchmark in order further understand the significance of the above frequencies. In absence of a directly relevant benchmark the findings of Hyland (2005b) may be of use. In a survey of published articles taken from 8 academic disciplines (including both the arts and sciences) Hyland reported the results contained in Table 3.

Stance Marker	Items Per 1000 Words	% of Total Use	Difference (% of Total Use: NIC and Hyland study)
Uncertainty Marker (Hedge)	14.5	0.59	+0.18
Certainty Marker (Booster)	5.8	0.24	-0.16
Attitude Marker	4.2	0.17	-0.02
Total	24.5	100	---

Table 3: Epistemic Stance Markers in Academic Articles (Hyland, 2005b).

Overall the authors of the NIC report were less likely to intrude upon the text to provide epistemic or affective commentary than the academic authors of the articles used in Hyland's study. Given the highly uncertain nature of the subject matter (i.e. the future) this was rather surprising as was the amount of propositional content presented in a bald manner i.e. no qualification.

When the choice to use stance was made the authors of the NIC report were significantly more likely than academic authors to use hedges (+18%) and less likely to use boosters (-16%) (Mann Whitney Test). In terms of the sub-categories, the authors used more hedges in relation to *Efficiency Gains* and *Military Investment.* Boosters were more often used in relation to the subcategory of *Technological Imperative*. Attitude markers were mainly used in relation to the subcategory of *Medical Improvements*.

5 Summary

The high level of macro-propositional agreement and the amount of material presented in a bald fashion suggests that the NIC took a largely unquestioned, techno-optimistic stance. The lack of consideration for radical thinking and possibilities is rather worrisome given the heavy burden placed on the planet by the current economic hegemon.

References

Commoner, B. 1971. The Closing Circle –Nature, Man, and Technology. New York: Aldred. A. Knopf.

Dickson, D. 1975. The Politics of Alternative Technology. New York: American Management Association.

Eagleton, T. 1991. Ideology. An Introduction. London: Verso.

Greening, L.A., Greene, D.L. and Difiglio, C. 2000. Energy Efficiency and consumption –The Rebound Effect-A Survey. Energy Policy 28 (3): 389-401.

Hyland, K. (2005a). Stance: Exploring Interaction in Writing. London: Continuum.

Hyland, K. (2005b). Stance and Engagement: A model of interaction in academic discourse. Discourse Studies 7(2): 132-191.

Huesemann, M. and Huesemann, J. 2011. TECHNO-FIX: Why Technology Won't Save Us or the Environment. Gabriola Island: New Society Publishers.

Jevons, W.S. 1865. The Coal Question –Can Britain Survive? London: Macmillan.

MacKellar, F.L. 1996. On Human Carrying Capacity –A Review Essay on Joel Cohen's 'How Many People Can the Earth Support'. Population and Development Review 22 (1): 145-156.

McKeown, T. 1979. The Role of Medicine – Dream, Mirage or Nemesis. Princeton: Princeton University Press.

Pollon, M. 2007. The Omnivore's Dilemma –A Natural History of Four Meals. New York: The Penguin Group.

Popper, K.R. 1957. The Poverty of Historicism. Routledge.

Runge, C.F.B., Senauer, P.G., Pardey, P.G. and Rosegrant, M.W. 2003. Ending Hunger in Our Lifetime –Food Security and Globalization. Baltimore: John Hopkins University Press.

Scharlemann, J.P.W., and Laurence, W.F. 2008. How Green Are Biofuels. Science, 319:43-44.

Sinclair, J. 1991. Corpus Concordance and Collocation. Oxford: Oxford University Press.

Taleb, N.N. The Black Swan: The Impact of the Highly Improbable. London: Penguin.

Wright, R. A Short History of Progress. Da Capo Press.

National Intelligence Council. 2004. Mapping the Global Future: Report on the National Intelligence Council's 2020 Project. Accessed here: http://www.futurebrief.com/project2020.pdf

Russian in the English mirror: (non)grammatical constructions in learner Russian

Evgeniya Smolovskaya
National Research University HSE
esmolovskaya@hse.ru

Evgeniy Mescheryakova
National Research University HSE
eimescheryakova@hse.ru

Olesya Kisselev
Pennsylvania State University
ovk103@psu.edu

Ekaterina Rakhilina
National Research University HSE
erakhilina@hse.ru

1 Introduction

Learner corpora have truly become an irreplaceable resource in the study of second language acquisition (SLA) and second language pedagogy in the recent decades. Although the majority of learner corpora to date represent English as a Foreign (FL) or Second (L2) language, many well-designed corpora of learner languages other than English have appeared in the past decade. A new linguistic resource, known as Russian Learner Corpus (RLC, http://web-corpora.net/heritage_corpus), is now available for researchers of L2 Russian. RLC is a collaborative project between the Heritage Russian Research Group (Higher School of Economics) under E. Rakhilina and a team of American researchers associated with the Heritage Language Institute (M. Polinsky, O. Kisselev, A. Alsufieva, I. Dubinina, and E. Dengub). The corpus includes comparable sub-corpora created by speakers of FL Russian and speakers of Russian as a Heritage language (HL), across different levels of language command, linguistic modes (written and oral) and genres. The new corpus provides a unique opportunity to conduct comparative studies in Russian SLA and pedagogy, as well as methodological studies that have relevance for learner corpora annotation, analysis and management.

2 Error analysis of Learner Russian

The idea of usefulness of error analysis has been largely -- if not uncritically -- embraced by the field of Learner Corpus research (Granger 1998). The main discussions vis a vis systematic errors in learner language are currently focusing on the following two issues: 1. methodological issues such as creating annotator-friendly tagging systems and automated and semi-automated methods of error identification in non-standard texts, and 2. theoretical issues of error identification, categorization and explanation of error source. These two lines of work are not entirely independent of each other; in fact, they feed into one another, ideally, resulting into creation of a unified, automated, and comprehensive error tagging system. Error analysis of the texts in the Russian Learner Corpus has been thus far attempted from these two perspectives. Klyachko et al. (2013) tested a protocol for automated error identification, which consisted of comparison of lists of bi- and tri-grams found in the learner corpus to the lists of bi- and tri-grams found in a native corpus. This approach was found to be fairly successful in identification of such errors as noun-adjective agreement and prepositional and verbal government. However, it comes with certain limitations: for instance, it provided far less accurate results for discontiguous structures compared to contiguous strings (possibly due to the size and characteristics of the baseline corpus) and, more importantly, left a large repertoire of non-grammatical structures out of its scope.

Another approach, discussed in this paper, begins with manual annotation of a sample of learner texts. The annotators first read and tag deviant forms using a tagging software developed for the project (see the illustration of the program interface below, Figure 1). Importantly, the error tags include the information about the source of an error (calque, semantic extension, etc.), in addition to the information about the structural property of an error (e.g. lexical, aspectual, morphological).

Those erroneous structures that reach a frequency threshold that reliably points to a systematic rather than a random nature of these errors are then examined and grouped according to structural and functional properties. To illustrate how this approach works we refer to examples below:

(1) * eto vredno svoim pal'cam
* *it is bad one's* DAT PL *fingers* DAT PL
cf. eto vredno dlya PREP pal'cev GEN PL
it is bad for fingers
**Но, по-моему, это вредно своим пальцам, поскольку часто встречающиеся буквы не находятся близко к центру клавиатуры*
(L2 speaker)
But I think it is bad for one's fingers since the most frequent letters are not located towards the center of the keyboard.

(2) * eto ne trudno govorit'
* *it is not hard to speak*
cf. NULL ne trudno govorit'
(it's) not hard to speak
**С этим человеком, это не трудно говорю, потому что мы понимаем друг друга.*
(L2 speaker)

With this person, it's not hard to speak because we know each other.

In analyzing errors like these, we attempt to establish those patterns and rules present in the interlanguage of the learner that allow us to hypothesize (and in some cases predict) the source of the non-native-like construction. Thus, in example 1, the likely source of error is the English (albeit infrequent) construction *to be bad (harmful)+to+something*. For instance:

(a) *Ayscough felt that white glass created an offensive glaring light that was bad to the eyes.* (GloWbE)

(b) *On the other hand, we may find out 3D is truly harmful to children's eyes, at which point it will likely lose the interest of the public and die.* (GloWbE)

The transfer is likely to be supported by the existence of two possible constructions in Russian as well, *dlya*(cf. for)+GEN and NULL PREPOSITION+DAT. These two constructions are close semantically and may be interchangeable (Ladygina 2014) under the right circumstances, i.e. if the experiencer is animate (Bonch-Osmolovskaya 2003). In example 1, the requirement of animacy is not upheld (likely because no such restraint exists in English). Interestingly, HL learners (at least at advanced levels) appear to be sensitive to the restraint of animacy and do not exhibit errors of this type.

Example 2 (ETO+AVD+INF) belongs to a type of learner errors known as null subject errors; it is frequently mentioned in the works on negative transfer. Although this error type is most often explained by the negative transfer from English, persistence of such errors in HL interlanguage indicates that it is also preempted by the fact that

Russian allows for pronoun *eto* in certain constructions, i.e. INF(COP)+ADV-o/ INF (COP) – *eto* +ADV-o:

(c) Купить в супермаркете пищу и из-за нее потом едва не протянуть ноги – это сейчас несложно. (Russian National Corpus)
To buy groceries in a supermarket and then almost die as a result – it's not difficult these days.

More importantly, the previous research in this area of grammar disregarded diachronic development of the use of *eto* in the Russian language. Thus in the main corpus of the Russian National Corpus, we find the following dynamic: in the text authored in the 19th century, the frequency of *eto*-construction is 1.4*10^-5, in the 20th century texts it becomes 2.87*10^-5, and in the texts authored in the first decade of the 21st century the frequency reaches 3.35*10^-5. Thus, the construction under examination has become 105.3% more frequent in the 20th century when compared to the 19th century, and 16.4% more frequent in the 21st when compared to the 20th.

(d) Думаешь, это было просто — бросить все и прилететь сюда? (Russian National Corpus)
You think it was easy – to drop everything and fly here?

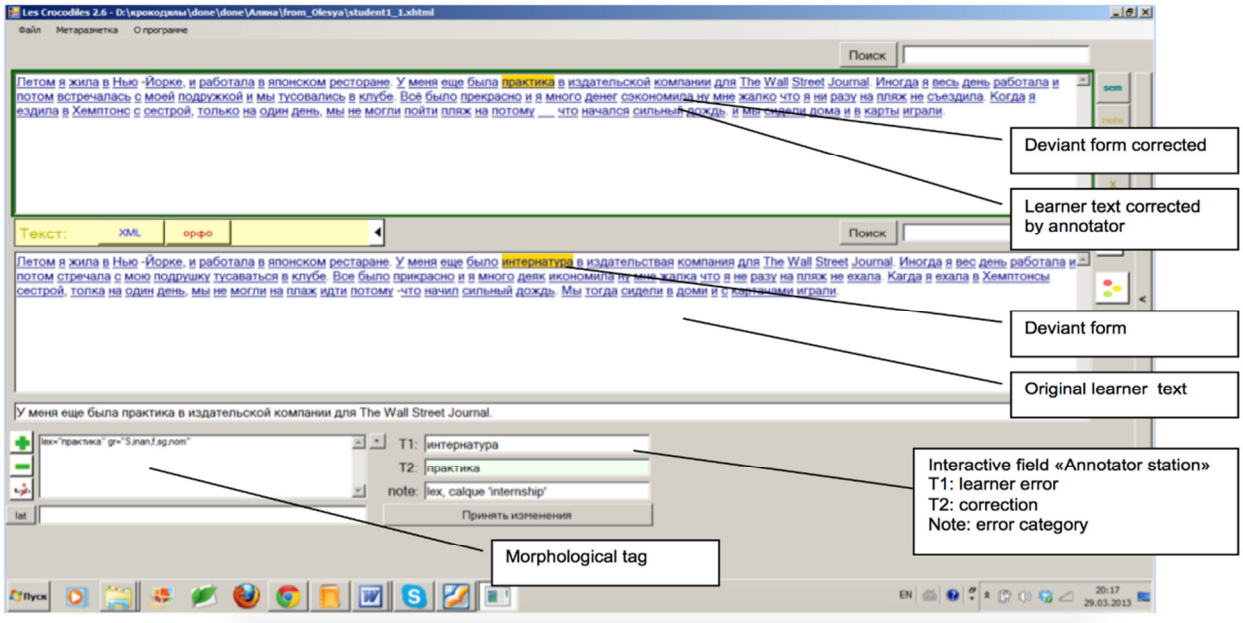
However, when it comes to the examination of oral sub-corpus of the RNC, we find the construction

ETO+ADV-o – INF(COP):

(e) Это тяжело очень сказать / когда достроят (Russian National Corpus)

It is hard to say / when they will finish building.

Additionally, in constructions that employ *kak* (cf.



how) the word order is the same as in English: *kak*+*eto*+ADV – INF (cf. Eng., how it is+ADJ+INF)

(f) … как же это сложно: говорить так, чтобы тебя слышали и слышали.

(Russian National Corpus)

how it is difficult – to speak in a way that you are listened to and heard.

In other words, the learners (and error-taggers) have to follow two sets of rules for *eto*-constructions: one in writing, another in speech.

Such error analysis is not methodologically simple: it requires extensive analysis of errors and comparable or similar constructions in the native and target language. However, we believe that this approach will allow us to build a detailed and comprehensive repertoire of error types and to build a library of error "models" (effectively represented by strings of morphological tags such as *eto*+ADV+INF). These models will be subsequently incorporated into a tagging software used to automatically detect and annotate errors in constructions in non-standard varieties of Russian.

3 Conclusions

The paper illustrates the general approach to the identification, categorization and explanation of errors in learner Russian. Although this approach comes with a list of challenges and limitations, we believe that it will not only significantly improve the Russian Learner Corpus but will provide a new model for error-annotation for other corpora "with noise in the signal".

References

Alsufieva, A., Kisselev, O. and Freels, S. 2012. "Results 2012: Using Flagship Data to Develop a Russian Learner Corpus of Academic Writing". *Russian Language Journal,* 62: 79-105.

Bonch-Osmolovskaya, A. 2006. Dativnyj subject v russkom yazyke: korpusnoe issledovanie. Unpublished PhD thesis, Moscow State University.

Granger, S. 1998. *Learner English on Computer*. Addison Wesley Longman, London and New York.

Ladygina, A. 2014. Russkie heritazhnye konstruktsii: korpusnoe issledovanie. Unpublished MA thesis, Moscow State University.

Klyachko, E., Arkchangelskiy, T., Kisselev, O. and Rakhilina. 2013 Automatic error detection in Russian learner language. Conference presentation, CL2013

Discourse and politics in Britain: politicians and the media on Europe

Denise Milizia

University of Bari "Aldo Moro"

denise.milizia@uniba.it

This research is part of a larger project that investigates the sentiment of the UK towards the European Union, the British "à la carte" attitude to the EU, this cherry-picking attitude, as it has been called, which sees Britain opting in, opting out, in many ways half in, half out (Musolff 2004). It cannot be denied that Britain has always been an awkward partner in EU affairs (George 1994), agreeing to some policy areas, disagreeing to some other European policies, for the sake of what has now become the signature of this government: 'in the national interest', 'in Britain's national interest' (Milizia 2014a).

This investigation is based on two political corpora, a spoken corpus and a written corpus. The spoken corpus includes all the speeches of the Labour government from 1997 to 2007, led by Tony Blair, and from 2007 to 2010, led by Gordon Brown; it also includes all the speeches of the coalition government formed in 2010, in which Conservative Prime Minister David Cameron and Liberal Democrat Deputy Prime Minister Nick Clegg were more often than not at odds over the position that the UK will have to take in the near future; it also includes some speeches of the current government, the Conservative government led by David Cameron, who is back in Downing Street after winning the general election of May 2015. Furthermore, the corpus includes some speeches delivered by Nigel Farage, former leader of UKIP (United Kingdom Independence Party), who wants the UK "unequivocally out of Europe", promising that "an exit is imminent" (Milizia 2014c), and some speeches by Ed Miliband, former Labour leader who, in the 2015 Manifesto, maintained that David Cameron "is sleepwalking Britain towards exit from the European Union", and that "Britain will be better off remaining at the heart of a reformed EU".

At the time of writing the spoken corpus totals slightly more than 5 million words.

The written corpus relies on articles from *The Economist*. The data selected comes from the section World Politics, Europe, and at the time of writing it counts 2 million words.

The purpose of the present investigation is to analyse and compare how British politicians and this élite magazine mediate specialized knowledge, European political knowledge in the case in point, how they disseminate knowledge and how they

transform 'exclusive expertise' into 'comprehensible knowledge', namely how popularization discourse is formulated for the non-specialised addressee to make it accessible to the lay public (Calsamiglia and van Dijk 2004).

The Economist may be defined as an élite magazine: it targets a highly educated readership, and presupposes some technical lay knowledge of its readers. Thus, in its pages some knowledge is presupposed, other kind of knowledge is reminded, and some other times knowledge is expressed and newly constructed. *The Economist* and British politicians' speeches are thus compared, to see if and to what extent the written and the spoken corpus share the same strategies for the management of discourse and politics.

The analysis carried out so far has shown that in both written and spoken discourse, in order to facilitate understanding of the main issues of the moment, several concepts are made accessible by metaphorization (Lakoff & Johnson 1980): the sink or swim metaphor, the slow lane/fast lane, the two-speed Europe, the one-way ticket, with the associated scenario of missing the EU train/bus/boat/ship/convoy have been in use for a while now. Indeed, the two-speed Europe is now being substituted with the multi-speed Union, and more recently, the European Union is being defined a "Teutonic Union", with Germany behind the scenes, quietly asserting its influence in Brussels. This paper focuses mainly on metaphors and analogies relating to the ill-fated relationship of the UK with the European Union and on whether "Britain and Europe can still save their marriage", which has indeed been on the rocks, as it were, for too long (Musolff 2000): Britain, in many ways, has been leaving the Union since virtually it became a member. Accordingly, the real question for British and Europe is whether the British will opt for a separation or for a divorce, even though an amicable divorce seems to many a pipe dream. Interestingly, in the coalition government, if the Prime Minister and the Deputy Prime Minister obviously agreed on certain matters, for example they were squarely decided that the UK will never give up the pound and join the euro, they did not see eye to eye on the referendum, which David Cameron had promised the British people if the Conservative party had won the general election. After the victory of the Tories at the national election, David Cameron will now have to deliver the in/out referendum on their future in Europe.

In this respect, Nick Clegg has often used the "play with fire" metaphor, saying that Cameron is playing with fire over UK European Union membership, "and if we go down this track, it is Britain that will get burned", becoming more marginalized, more isolated, regretting to be no longer part of the club (Milizia 2014a). Interestingly, David Cameron has used the same analogy in reference to the danger of staying in/out of the euro currency, voicing his misgivings that the Eurozone members are playing with fire with their plans to lock themselves into a United States of Europe (Semino 2002).

In an article of November 2013 titled "Little England or Great Britain?" *The Economist* depicts Britain as facing a choice between becoming more inward-looking and with less clout in the world or more outward-looking and surer of its identity and its place in the world. A few months later, in the two-day debate with Nigel Farage held before the European election last May, Nick Clegg borrowed the same words saying, "I want us to be Great Britain, not Little England".

The software used to process and compare data is *WordSmith Tools* 6.0 (Scott 2012).

References

Calsamiglia, H. and van Dijk T.A. 2004. "Popularization Discourse and Knowledge about the Genome". *Discourse & Society* 15 (4), Special issue *Genetic and genomic discourses at the dawn of the 21st century*, guest-edited by B. Nerlich, R. Dingwall, P. Martin: 369-389.

George, S. 1994. *An Awkward Partner. Britain in the European Community*. Oxford University Press:

Lakoff, G. and Johnson, M. 1980. *Metaphors we live by*. Chicago: Chicago Press.

Milizia, D. 2014a. "In, out, or half way? The European attitude in the speeches of British leaders". *Lingue e Linguaggi* 11: 157-175.

Milizia, D. 2014b. "Specialized discourse vs popularized discourse: the UK and the European Union". Paper presented at the University of Catania, Italy, 2nd International Conference, Language and Diversity: Discourse and Translation, 9-11 October.

Milizia, D. 2014c. "A bilingual comparable analysis: the European Union in the speeches of British and Italian leaders". Paper presented at the University of Milan, Italy, CLAVIER 14: LSP 20-21 November.

Musolff, A. 2004. *Metaphor and Political Discourse*. New York: Palgrave Macmillan.

Musolff, A. 2000. "Political imagery of Europe: A house without exit doors?" *Journal of Multilingual and Multicultural Development* 21 (3): 216-229.

Scott, M. 2012. *WordSmith Tools* 6.0. Lexical Analysis Software Limited.

Semino, E. 2002. "A sturdy baby or a derailing train? Metaphorical representations of the euro in British and Italian newspapers". *Text* 22 (1): 107-139.

Investigating the stylistic relevance of adjective and verb simile markers

Suzanne Mpouli
UPMC, LIP6, Paris
mpouli@acasa.lip6.fr

Jean-Gabriel Ganascia
UPMC, LIP6, Paris
jean-gabriel.ganascia@lip6.fr

1 Introduction

Similes are figures of speech in which the similarities as well as the differences between two or more semantically unrelated entities are expressed by means of a linguistic unit. This unit, also called marker, can either be a morpheme, a word or a phrase. Since similes rely on comparison, they occur in several languages of the world. Depending on the marker used and of the semantic or structural nuances it introduces, two main simile classifications have been proposed. From her study of images built on figures of resemblance in Baudelaire's poetry, Bouverot (1969) distinguishes between type I and type II similes. Whereas type I similes are denoted by a finite number of comparatives, prepositions or conjunctions which traditionally convey a comparison, type II similes are observed after a verb or an adjective which semantically contains the idea of similitude or difference. Leech and Short (2007) propose a stricter distinction and separate conventional similes of the form "X is like Y" from quasi-similes which revolve around all other linguistic constructions. This partition seems to take into account the fact that some simile markers are often preferred in a particular language. For example, 'like' in English or 'comme' in French are often presented as the prototypical simile markers and are more frequently used than the other markers.

Similes play an important role in literary texts not only as rhetorical devices and as figures of speech but also because of their evocative power, their aptness for description and the relative ease with which they can be combined with other figures of speech (Israel et al. 2004). Detecting all types of simile constructions in a particular text therefore seems crucial when analysing the style of an author. Few research studies however have been dedicated to the study of less prominent simile markers in fictional prose and their relevance for stylistic studies. The present paper studies the frequency of adjective and verbs simile markers in a corpus of British and French novels in order to determine which ones are really informative and worth including in a stylistic analysis. Furthermore, are those adjectives and verb simile markers used differently in both languages?

2 Adjective and verb simile markers

Comparison being a semantic construction, there exist no comprehensive and finite lists of verb and adjective simile markers. The choice of the adjective and verb simile markers used in this experiment has been based on their ability to introduce phrasal similes, i.e. similes in which the compared term is a common noun. As far as verb markers are concerned, were ignored impersonal constructions which only accept indefinite pronouns as subjects such as in the French sentence "J'aime sa voix, on eût dit une pluie de grelots".

Under the category of verb simile markers, were included modified forms not found in the literature such as 'be/become….kind/sort/type of' and 'verb+ less/more than' as well as their respective French equivalents 'être/devenir….espèce/type/genre/sorte de' and 'verb+ plus/moins que'. As a matter of fact, even though expressions such as 'kind of' or 'sort of' are often cited among metaphorical markers (Goatly 1997), they were judged not specific enough to explicitly introduce a comparison on their own. In the case of the comparison markers 'less than' and 'more than', the issue rather lies in the fact that they are also used in other types of constructions and conjoining them to a verb form seems a good compromise to restrict their polysemy.

In English, in addition to adjectives conveying the idea of comparison, adjective simile markers include adjectives formed by adding the suffix '-like' and by joining a noun to an adjective of colour. Those two types of adjectives are particularly interesting stylistically speaking because they can potentially be used to create neologisms. Table 1 lists all the adjective and simile markers used for the experiment.

	Verbs	**Adjectives**
English	*resemble, remind, compare, seem, verb + less than, verb + more than, be/become… kind/sort/type of*	*similar to, akin to, identical to, analogous to, comparable to, compared to reminiscent of, -like , noun+colour*
French	*ressembler à, sembler, , rappeler, faire l'effet de, faire penser à, faire songer à, donner l'impression de, avoir l'air de, verb + plus que, verb + moins que, être/devenir…espèce/type/genre/sorte de*	*identique à, tel, semblable à, pareil à, similaire à, analogue à, égal à, comparable à*

Table 1. Adjective and verb simile markers

3 Corpus building and extraction method

To extract simile candidates, digital versions of

various literary texts were collected mainly from the Project Gutenberg website [87] and from the Bibliothèque électronique du Québec[88,] Most of the novels included in the corpus were written during the 19th century so as to ensure linguistic homogeneity and because that century witnessed the novel imposing itself as a predominant literary genre. By observing a ration of least 3 novels per writer, we were able to put together a corpus of 1191 British novels authored by 62 writers and a corpus of 746 French penned by 57 novelists

Each corpus was tagged and chunked using TreeTagger, a multilingual decision tree part-of-speech tagger and a syntactic chunker (Schmid, 1994). The output of the tagger was further used to determine sentence boundaries. Each extracted sentence is considered a simile candidate if it contains one of the marker immediately followed by a noun-headed noun phrase.

4 Results

Since similes are realised first and foremost within a sentence, simile frequency was first calculated by dividing the number of occurrences of each simile marker in a particular novel by the total number of sentences in that novel. That measure however proves itself to not accurately reflect the distribution of verb and adjective simile markers as their use seems to not be influenced by the length of the novel. The frequency of each simile marker concerned in this experiment is therefore measured only by counting its occurrence in a specific novel.

Due to their polysemous use and the noise they introduce in the generated data, markers such as 'remind of' and 'rappeler' were not considered in the final analysis.

The discrepancy between the frequency count of each grammatical category in English and French tends to suggest that both languages work differently as far as simile creation is concerned. In English, verb simile markers appear to be more used than adjective ones. As a matter of fact, two main verb constructions largely surpass the other markers in number: the structures 'seem' + NP and 'be/become a sort/type/kind of...' + NP which count more than 5,000 occurrences.

Excluding adjectives formed by using '-like' or an adjective of colour as a suffix, 'akin (to)' is the most used adjective simile marker with about 350 occurrences and is generally associated with nouns denoting feelings.

As far as French is concerned, the gap between the use of adjective and verb simile markers is less pronounced. Like its English counterpart, 'sembler' has a good place among most frequently used verb markers but is less predominant than 'ressembler (à)'. French writers also distinguish themselves by their preference for introducing similes with the adjectives 'pareil (à)' and 'semblable (à)' .

In addition, in the French corpus, 'similaire (à)' is never used as a typical simile marker, just like 'identical (to)' in the British corpus. If 'identique (to)' appears in the French corpus, it has the smallest number of occurrences of all adjectives that are effectively used to create similes. Similarly, the inflected forms of 'comparable' are found in both corpora but are not so common. A possible explanation as Bouverot (1969) suggests could be the length of the word 'comparable' which surely affects the sentence rhythm.

From the results of this experiment, it is possible to conclude that there seems to exist preferred verb and adjective simile markers in fictional prose. However, even though some writers use them systematically—for example all texts by Balzac present in the corpus contain at least one occurrence of 'pareil à + NP'—the frequency use of these markers in those authors' novels generally vary greatly from one text to another. Consequently, even though verb and adjective simile markers could be interesting stylistic indicators, taken individually, they do not seem to be able to characterise unequivocally one author's style but rather hint at a possible group tendency or an aesthetic idiosyncrasy.

Acknowledgement

This work was supported by French state funds managed the ANR within the Investissements d'Avenir programme under the reference ANR-11-IDEX-0004-02.

References

Bouverot, D. 1969. "Comparaison et métaphore". *Le Français Moderne*. 37(2) :132-147.

Israel, M., Riddle Harding, J. and Tobin, V. 2004. "On Simile". *Language, Culture and Mind*. Stanford: CSLI Publications.

Goatly, A. 1997. *The Language of Metaphors*. London and New York: Routledge.

Leech, G. and Short, M. 2007. *Style in Fiction: A Linguistic Introduction to English Fictional Prose.* Harlow: Pearson Longman.

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. Proceedings of the International Conference on New Methods in Language Processing:

[87] www.gutenberg.org

[88] beq.ebooksgratuits.com

Competition between accuracy and complexity in the L2 development of the English article system: A learner corpus study

Akira Murakami
University of Birmingham
a.murakami@bham.ac.uk

1 Background

A recent trend in second language acquisition research is the investigation of learner language from three aspects; complexity, accuracy, and fluency, collectively referred to as CAF (Housen et al 2012b). Previous CAF studies have largely investigated the relationship of multiple features in second language (L2) development at the level of essay. That is, they tend to analyse what happens to a linguistic feature (e.g., type-token ratio) when the value of another feature (e.g., sentence length) changes across essays.

To date, however, there has been little work that investigated the trade-off between features at a finer level. The present study fills the gap. More specifically, the study focuses on the accuracy of the English article system and addresses the following research question: Is the accuracy of the L2 English article in the trade-off relationship with lexical and phrasal complexity at the sub-essay level?

2 Corpus

The study employed EF-Cambridge Open Language Database (EFCAMDAT). The corpus includes learners' writings at Englishtown, the online school run by Education First. A course in Englishtown consists of 16 Levels, each of which covers eight Units. Although learners are free to go back or skip units, they usually progress from lower to higher levels unit by unit. A placement test can suggest an appropriate level at which learners start their courses. Each unit includes a range of activities including receptive (listening and reading) and productive (speaking and writing) practice, as well as explicit grammar instruction on such features as articles and verbal inflection, and at the end of each unit is a free composition task on a variety of topics (e.g., self introduction, making requests). An example writing is provided at each composition task, and learners can freely refer to it during composition. They can also consult external resources like dictionaries in their writing. Each composition task specifies the length of the writing, which ranges from 20-40 words in Level 1 Unit 1 to 150-180 words in Level 16 Unit 8. Each writing receives manual feedback from teachers that includes the correction of erroneous article uses. The present study used the teachers' feedback as error tags and collected necessary information to calculate accuracy by exploiting them. Error tags are not annotated to all the writings, however. Apart from learners' writings, EFCAMDAT includes for each essay such metadata as the ID of the learner, his/her country of residence, the topic of the essay, the date and time of submission, and the Level and the Unit number for which the essay was written. This allows researchers to track the longitudinal development of individual learners. EFCAMDAT can, therefore, be viewed as a partially error-tagged longitudinal learner corpus.

The subcorpus used in the study included typologically diverse following 10 L1 groups; Brazilian Portuguese, Mandarin Chinese, German, French, Italian, Japanese, Korean, Russian, Spanish, and Turkish. Since EFCAMDAT does not provide us with the direct information of learners' L1s, they were inferred from the countries they reside as a close approximation. L1 Brazilian Portuguese, German, French, Italian, Korean, Russian, and Turkish learners correspond to those living in Brazil, Germany, France, Italy, Korea, Russia, and Turkey respectively. L1 Mandarin Chinese learners included those living in Mainland China or Taiwan, and L1 Spanish learners included those living in Spain or Mexico. The L1 groups were then divided into two groups. What I call the ABSENT group are those whose L1s lack the equivalent feature to English articles and include L1 Chinese, Japanese, Korean, Russian, and Turkish learners. The PRESENT group are the learners whose L1s have the equivalent feature and includes L1 Brazilian Portuguese, German, French, Italian, and Spanish learners. This variable is referred to as L1 type.

For the sake of data reliability, the present study only targeted the learners with at least 20 uses or errors of articles. The subcorpus consisted of approximately 67,800 essays by 10,800 learners, totalling six million words.

3 Analysis

The study used two measures of complexity; the number of words in the noun phrase (NP) in which an article occurs and the Guiraud's index calculated by dividing the number of types by the square root of the number of tokens. The former represents phrasal complexity, while the latter is a variant of type-token ratio and represents lexical complexity (House et al 2012a).

For each of the 466,800 uses and errors of the article, the following information was collected; (i) the NP length, (ii) the Gruiraud's index of the essay, (iii) the total number of words in the essay, (iv) the

overall proficiency of the learner represented by the average level of the online school the learner engaged in, (v) the number of essays the learner has written up to the point, and (vi) the learner's L1 type. The study built a mixed-effects binary logistic regression model (Barr 2008; Jaeger 2008; cf. Gries in press) that predicts the accuracy of each article use as a function of the six features above and their two-way interactions, and with the learner as the random-effects factor.

4 Results and Discussion

The study observed a clear trade-off between accuracy and complexity. The main findings include the following:

- The longer the NP, the more likely the use of the article is erroneous.
- An interaction is observed between NP length and proficiency. Interestingly, while proficient learners generally outperform less proficient learners, those of higher proficiency are more strongly affected by NP length. In other words, the effect of NP length is not only unmitigated but also enhanced in higher proficiency learners.
- L1 type does not mitigate the effect of NP length. The ABSENT and the PRESENT groups are equally influenced by phrasal complexity.
- The higher the Guiraud's index, the more likely the learner makes an error in the article use.
- The effect, however, is not observed in the learners of higher proficiency. Higher proficiency learners can achieve both high lexical complexity and article accuracy.
- The effect is also mitigated by L1 type. The PRESENT group is less likely to be affected by lexical complexity.
- The findings above are subject to individual variation. The magnitude of individual variation, however, varies across the variables. While individual variation easily reverses the effect of Guiraud's index, the variation is much smaller in the effect of NP length. In other words, although a number of learners achieve higher article accuracy in the writings with higher lexical complexity, much fewer mark higher article accuracy in longer NPs.

The last point is illustrated in Figure 1 and Figure 2. Figure 1 shows the predicted accuracy in varying NP length across individual learners. The horizontal axis represents the number of words in the NP excluding the article and the noun described by the article, or more simply put, the number of intervening words between the article and the noun. Zero indicates that there is no word in between (e.g., a book), while one indicates the presence of one word (e.g., an interesting book), and so forth. The vertical axis represents accuracy or the probability that the learner correctly uses the article. Each line represents the predicted accuracy of one learner at the mean proficiency level and the mean lexical complexity. The left panel includes the ABSENT learners and the right panel includes the PRESENT learners. We can see from the figure that, while individual variation is present in the effect of NP length on accuracy, there is a strong overall tendency that the accuracy decreases in longer NPs (i.e., most of the lines decrease from left to right).

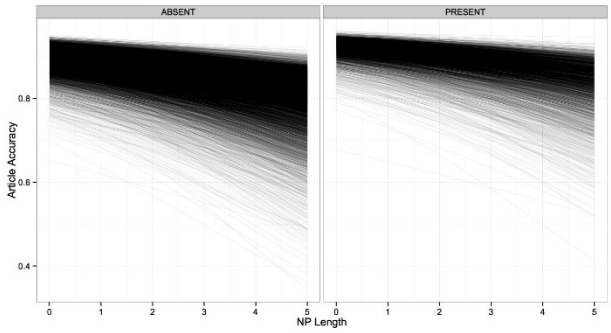


Figure 1. Individual variation in the effect of NP length on article accuracy across L1 types

Figure 2 is similar to Figure 1, except that the horizontal axis now represents Guiraud's index. Unlike the effect of NP length in Figure 1, the figure clearly demonstrates large individual variation in the effect of Guiraud's index. While there is still a tendency that article accuracy tends to be lower when the value of Guiraud's index is high, this tendency is easily outweighed by individual variation. There are a nontrivial number of learners whose article accuracy is lower in the writings with lower values of Guiraud's index.

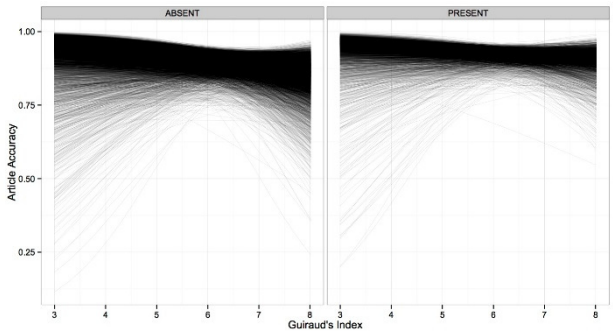


Figure 2. Individual variation in the effect of Guiraud's index on article accuracy across L1 types

The findings suggest the competition of cognitive resources between accuracy and complexity at a local level and also point to the complex nature of

L2 accuracy development.

References

Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474. doi:10.1016/j.jml.2007.09.002

Gries, S. T. (in press). The most underused statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora*. Retrieved from http://www.linguistics.ucsb.edu/faculty/stgries/research/ToApp_STG_MultilevelModelingInCorpLing_Corpora.pdf

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. doi:10.1016/j.jml.2007.11.007

Housen, A., Kuiken, F., & Vedder, I. (2012a). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 1–20). Amsterdam: John Benjamins.

Housen, A., Kuiken, F., & Vedder, I. (2012b). *Dimensions of L2 performance and proficiency: Compmlexity, accuracy and fluency in SLA*. Amsterdam: John Benjamins.

Metaphor in L1-L2 novice translations

Susan Nacey
Hedmark University College
susan.nacey@hihm.no

1 Introduction

Metaphor is frequently viewed as a translation problem, "a kind of ultimate test of any theory of translation" (Toury 1995: 81). Much focus in translation studies has revolved around the extent to which metaphor is translatable and the development of guidelines for metaphor translation. In addition, a growing body of research is being produced in the field of Descriptive Translation Studies, investigating what translations actually are rather than what they should be (see e.g. Rosa 2010; Toury 1995). The present study contributes to this endeavor through an exploration of the translation of metaphors found in the Norwegian-English Student Translation Corpus (NEST),[89] a corpus of L2 learner language.

This investigation identifies and categorizes the translation of metaphors from 30 different Norwegian source texts (ST) in a total of 287 English translated texts (TT), thereby both describing individual translations and providing comparative descriptions of several TTs derived from the same ST. The paper focuses on the translations of three types of metaphors, identified using MIPVU (cf. Steen et al. 2010): 1) metaphorical verbs, codified in Norwegian, 2) idioms, which are often culture-specific and 3) potentially deliberate metaphorical expressions such as similes and other figurative analogies (cf. Nacey 2013: 168-173; Steen, 2008).

2 Informants and material

The informants have Norwegian as their L1, and are advanced L2 learners of the target language (TL) enrolled on one of several parallel tertiary-level translation courses taken as part of their English studies. The main goal of the translation courses was to raise language awareness, thereby increasing these learners' English proficiency. Particular focus was placed upon Norwegian-English contrastive differences, both in terms of language and style. Some theory of translation was nonetheless included in the courses, even though they did not qualify students as translators. Of particular relevance for the current study was the emphasis placed on the principle of faithfulness, maintaining that the TTs

[89] Nest is found here: http://clu.uni.no/humfak/nest/.

should mirror the STs as closely as possible. The TTs are thus 'overt translations' inextricably and explicitly linked to the STs rather than 'covert translations' intended as original texts in their own right for their target audience (a distinction made by House 2010).

The STs range from 200 to 900 words and cover many different topics and text types, so as to illustrate a variety of contrastive challenges for the learners to translate and discuss. Texts thus range from instructional pamphlets and newspaper articles to fictional prose. Most STs have multiple translations (mean = 9.5 TTs per ST).

3 Categorization of metaphor translations

Translated metaphors have been categorized following a version of Newmark's (198: 88-91) proposed guidelines for translating metaphor, listed in Table 4 along with the translation strategy abbreviations adopted in this paper.

	Translation strategy	Abbreviation
1	Reproduction of the same SL metaphor in the TL	M → M
2	Replacement of the SL metaphor with a standard TL metaphor	$M_1 \rightarrow M_2$
3	Translation of the SL metaphor by simile	M → S
4	Translation of metaphor (or simile) by simile plus sense [a literal gloss]	M/S → S + gloss
5	Conversion of metaphor to sense [a literal paraphrase]	M → P
6	Deletion of metaphor	M → Ø
7	Translation of metaphor by same TL metaphor plus sense [a gloss]	M → M + gloss

Table 4. Classification guidelines for metaphor translation

Newmark's proposed procedures constitute a top-down approach, based on an assumption that the translators want to render the metaphors "as accurately as possible, not to pare them down" (Newmark 1981: 87). Actual translation occurrences were not consulted when drawing up the guidelines (see Fernández 2011: 265). Thus, the present study adapts Newmark's classification system, modifying it as indicated by the translation solutions actually chosen by the students – thereby ending up with a classification that represents the data under study.

4 Sample analysis: Idioms

The NEST STs contain relatively few idioms, not unsurprising given e.g. Moon's (2007: 1050) research indicating that smaller corpora (< 100 million words) yield only isolated instances of idioms, except for 'anomalous local densities' of an idiom repeated in a single text. Nevertheless, because comprehension of unfamiliar idioms often depends upon some degree of shared cultural knowledge, they are of interest when investigating translation strategies of metaphor. Translation of idioms may pose particular problems when it comes to the balance between faithfulness to the ST and production of a TT that is both understandable and idiomatic for the text type in question. One NEST idiom is found in a ST about the life of Norwegian author Bjørnstjerne Bjørnson. He is described as being an independent individualist with a characteristic 'kjerringa-mot-strømmen-holdning' [literal: hag-against-stream-attitude]. The phrase derives from a Norwegian folktale tale where a disagreeable wife argues with her husband about the best way to harvest grain. While he intends to mow the grain with a scythe, she insists that it be cut with shears; the husband finally silences his wife's nagging by drowning her in a nearby river. He later searches for her body to give her a proper funeral, only to find that she has drifted upstream, against the current. The idiom thus refers to people who are both stubborn and irritating, who do what they want without listening to others. While variants of this folktale are known in other cultures, there is no traditional English equivalent. Packing so much cultural information into a comprehensible English translation is challenging for novice translators, ten of whom translated this text. Their solutions are presented in Table 5 showing the different translations, NEST tags identifying TT and sentence, and categorization of translation strategy.

Only a single student chose an approximate literal paraphrase (M → P), this being the least popular translation strategy. Although all the others retained metaphor, none chose a pure M → M approach, with a literal transliteration of each element of the idiom. They have thus realized that an English readership may not have the necessary cultural background knowledge to fully understand the phrase when rendered word-for-word, and have produced alternative versions. In most cases, 'kjærringa' (literal: hag) has been dropped in the English version (hence the minus symbol in the translation strategy code). The one exception is Translation 2, where the core elements of the phrase remain in the original Norwegian (presumably evaluated as untranslatable), followed by lengthy explicitation – making this version arguably the least idiomatic of the ten translations. Six of the nine remaining cases retain the image of resistance to flowing water, alternatively translated as 'stream' (influenced by the partial false friend in the ST, Norwegian 'strøm'), 'current', or 'currant' (a spelling error). Two of these six add information to the metaphor by introducing the element of swimming (hence the plus symbol in the translation strategy code);

swimming is, however, incoherent with the original story as the wife had been drowned, meaning that her body floated rather than swam.

	Translation	TT ID tag (NEST_Opp_)	Translation strategy
1	characteristic for his "against-the-stream-attitude"	002en.s32	M → M (-)
2	characteristic to his "kjærringa mot strømmen" attitude (the Norwegian folktale about the old woman who always had to have her own way")	003en.s37	M → M (L1) + gloss
3	characteristic for his "going against the grain attitude"	004en.s37	$M_1 \rightarrow M_2$
4	typical of his go against the stream-attitude	005en.s37	M → M (-)
5	characteristic for his "go against the grain" attitude	007en.s37	$M_1 \rightarrow M_2$
6	characteristic of his "swimming upstream-nature "	008en.s30	M → M (-/+)
7	characteristic for his go against the grain-attitude	010en.s36	$M_1 \rightarrow M_2$
8	characteristic for his "swimming-against-t(Steen et al., 2010)he-currant-attitude"	011en.s39	M → M (-/+)
9	characteristic of him to go against the current	014en.s43	M → M (-)
10	characteristic for his attitude of contrariness	159en.s38	M → P

Table 5. Translations of 'karakteristisk for hans kjerringa-mot-strømmen-holdning' (NEST_Oppno.s38)

Three of the students chose to substitute another TL metaphor, 'go against the grain', for the SL metaphor, the $M_1 \rightarrow M_2$ strategy. The two metaphors are close in terms of semantics, but the M_2 metaphor introduces certain connotations absent from the SL metaphor – that is, someone doing something against the grain is performing an action unexpected of them contrary to one's normal inclination or disposition. By contrast, the wife from the folktale behaves true to form.

These translations offer several indications that the informants are still very much English language learners – this may be noted by the choice of 'stream' where 'current' or 'flow' might be more appropriate, by the spelling error 'currant', and by the apparent lack of realization of the added connotation of the M_2 metaphor. In addition, most of the students demonstrate colligation problems, not realizing the standard English colligation is 'characteristic of'. The most common choice of preposition is 'for', the basic translation of Norwegian 'for' that is appropriate for the SL context. Nevertheless, what is evident from these translations is that all the informants in some way acknowledged the translation challenge raised by this idiom, by attempting to unpack the SL metaphor and repack it in the TL.

References

Fernández, E. S. 2011. "Translation studies and the cognitive theory of metaphor". *Review of Cognitive Linguistics,* 9 (1), 262-279.

House, J. 2010. Overt and covert translation. In Y. Gambier & L. Doorslaer (eds.), *Handbook of Translation Studies, Volume 1* (pp. 245-246). Amsterdam: John Benjamins.

Moon, R. 2007. "Corpus linguistic approaches with English corpora". In H. Burger (ed.), *Phraseologie: Ein internationales Handbuch der zeitgenössischen Forschung* (Vol. Halbbd. 2, pp. 1045-1059). Berlin: Mouton de Gruyter.

Nacey, S. 2013. *Metaphors in learner English*. Amsterdam: John Benjamins.

Newmark, P. 1981. *Approaches to translation*. Oxford: Pergamon Press.

Rosa, A. A. 2010. "Descriptive Translation Studies (DTS)". In Y. Gambier & L. Doorslaer (eds.), *Handbook of Translation Studies, Volume 1* (pp. 94-104). Amsterdam: John Benjamins.

Steen, G. J. 2008. "The Paradox of Metaphor: Why We Need a Three-Dimensional Model of Metaphor". *Metaphor & Symbol, 23*(4), 213-241.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., Krennmayr, T., & Pasma, T. 2010. *A method for linguistic metaphor identification: from MIP to MIPVU*. Amsterdam: John Benjamins.

Toury, G. 1995. *Descriptive translation studies - and beyond (Vol. 4)*. Amsterdam: John Benjamins.

Effects of a Writing Prompt on L2 Learners' Essays

Masumi Narita
Tokyo International University
mnarita@tiu.ac.jp

Mariko Abe
Chuo University
abe.127@g.chuo-u.ac.jp

Yuichiro Kobayashi
Toyo University
kobayashi0721@gmail.com

1 Introduction

Since the development of computerized learner corpora has flourished, extensive research has been conducted on linguistic features that characterize second language (L2) learners (Granger 1998; Tono et al. 2012; Granger et al. 2013; Ishikawa 2013). Within this line of research, we explored similarities and differences in linguistic features used in argumentative essays produced by English language learners in four Asian countries (Hong Kong, Taiwan, Korea, and Japan) (Abe et al. 2013). One of our major findings is that lexical words in a given essay prompt tended to be used repetitiously by these L2 learners, which is consistent with the results found by Liu and Braine (2005).

As proposed by Halliday and Hasan (1976), lexical repetition as a cohesive device plays an important role in creating textual coherence. To look further into lexical cohesion in our learner writing, we need to explore the use of multi-word sequences (referred to as lexical bundles or N-grams) as well as individual word forms. In the present study, therefore, the following research questions are addressed using the same learner data analyzed in our previous study.

1) Given the same writing task, how do L2 learners in four Asian countries use lexical bundles, particularly those affected by the writing prompt?

2) What is the relationship between the use of prompt-affected lexical bundles and the learners' English proficiency?

2 Research Methodology

We used four sub-corpora of the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa 2013). The four learner corpora consisted of argumentative essays produced by L2 learners (college students) in four Asian countries in response to the writing prompt, "*It is important for college students to have a part-time job.*" Since English proficiency based on the Common European Framework of Reference for Languages (CEFR) is also available in the ICNALE data, learners' lexical use can be compared according to the difference in their English proficiency as well as their home country. The ICNALE also provides essays produced by native speakers of English within the same task conditions; therefore, we can briefly sketch the present L2 learners' lexical use with reference to the comparable data.

An N-gram analysis tool was newly developed to compute the frequency of N-grams while preventing duplicated counting. In the present analysis, N-gram patterns were generated from the writing prompt only, and the value of N was set to 3 (minimum) through 11 (maximum; i.e., the total number of words in the given writing prompt). When all the N-gram patterns and their respective frequencies were extracted from each essay, the usage ratio of the N-gram wordings to essay length was calculated.

3 Results

The usage (or matched) ratio represented as a percentage for each learner group is illustrated in Figure 1, where the usage ratio in the essays produced by native speakers of English is also included. [90] The median usage ratio of N-gram wordings to essay length for each group is shown by a thick solid line within each box. The dotted horizontal line represents the overall median usage ratio.

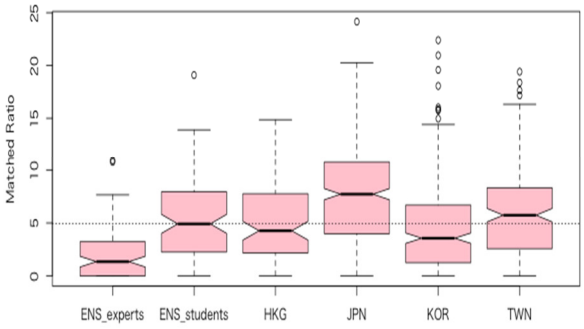


Figure 1: Usage (matched) ratio (%) of N-gram wordings to essay length

The notched box plots in Figure 1 indicate 1) that Japanese learners of English used more prompt wordings than the other learner groups, 2) that Taiwanese and particularly Hong Kong learners used prompt wordings to almost the same extent as native English-speaking college students, and 3) all

[90] The comparable essay data produced by native speakers of English can be broken down into two groups: 100 ENS_experts (employed adults such as English instructors, translators, and company employees) and 100 ENS_students (college students).

the student groups used far more prompt wordings than native English-speaking experts.

Table 1 shows the median usage ratio of prompt wordings according to English proficiency in each learner group. Leaving aside the C1_0 level learners due to a limited number of essay data, it is evident that the three learner groups except Hong Kong learners tended to use more prompt wordings the higher their proficiency level. This finding is contrary to our expectation that less proficient learners would be more affected by the prompt in their L2 writing than more proficient peers.

CEFR Level	Hong Kong	Taiwan	Korea	Japan
A2_0	5.05	4.78	2.97	7.91
B1_1	4.19	5.02	3.25	7.49
B1_2	3.85	6.34	3.24	8.54
B2_0	5.57	7.29	4.94	8.06
C1_0	5.61	7.81	4.39	6.19

Note: The number of C1_0 level learners is 2 in Hong Kong, 18 in Korea, and only 1 in both Taiwan and Japan.

Table1: Median usage ratio (%) according to L2 learners' English proficiency

4 Discussion

The major findings described in the previous section can possibly be explained by our observations: 1) the variation of prompt-affected lexical bundle use increased across the learners' English proficiency levels, particularly across A2_0, B1_1, and B1_2 levels, and 2) the prompt sentence (i.e., the whole wording of the prompt) was persistently used by Japanese, Taiwanese, and Korean L2 learners regardless of proficiency levels, whereas not a single instance was found in the essays produced by Hong Kong L2 learners. More proficient L2 learners in the present study appeared to elaborate on their argumentation using more prompt-induced word sequences. Furthermore, our analysis revealed that Japanese L2 learners tended to use the whole prompt wording both at the beginning and ending of their essays more frequently than their Taiwanese or Korean counterparts.

The location and rhetorical role of a given prompt sentence reused or recycled in L2 writing suggest an L2 learners' writing strategy to produce more language that has been developed through the writing instruction they received. Learners in the present study were most likely to reuse the given prompt sentence at the beginning of their essays so that they could clarify their standpoint and determine an overall direction of their argumentation. More interestingly, L2 learners frequently reused the prompt sentence twice, both at the beginning and ending of their essays, to structure their essays according to the basic organizational scheme they have learned through formal instruction in the classroom. The absence of reusing the prompt sentence in Hong Kong learners may reflect the differences in L2 learning environments and/or L2 writing instruction.

It is interesting to note that this prompt sentence recycling was also observed among native English speakers: 15 instances in the essays produced by college students and 6 instances in those produced by experts. This observation may provide insight into developmental aspects of argumentative essay structuring.

5 Conclusion

The present study examined the use of prompt-affected lexical bundles in argumentative essays produced by four Asian learner groups. Our quantitative analyses revealed that among these learner groups, more recycling of prompt wordings was found in the essays produced by Japanese L2 learners. Also, unlike the other learner groups, Hong Kong L2 learners did not reuse the whole wording of the prompt in their essays at all. Furthermore, it was found that L2 learners tended to reuse parts or the whole wording of the given writing prompt regardless of their English proficiency.

We assume that recurrent use of lexical bundles, whether or not prompt-induced ones, is relevant to the development of lexical cohesion in L2 writing. Thus, further corpus-based research is necessary to obtain a better understanding of lexical cohesive links produced by L2 learners. Further research can also offer pedagogical implications for L2 writing instructors to develop more informed courses and materials, enabling their students to produce more cohesive written discourse in a second language.

Acknowledgements

This research was supported by Grants-in-Aid for Scientific Research Grant Numbers 24320101 and 26370703.

References

Abe, M., Kobayashi, Y. and Narita, M. 2013. "Using multivariate statistical techniques to analyze the writing of East Asian learners of English". In S. Ishikawa (ed.), *Learner corpus studies in Asia and the world - Vol.1*. Kobe: School of Language and Communication, Kobe University.

Granger, S. 1998. *Learner English on computer*. New York: Addison Wesley Longman, Inc.

Granger, S., Gilquin, G. and Meunier, F. (eds.) 2013. *Twenty years of learner corpus research - looking back, moving ahead: Proceedings of the first learner corpus research conference (LCR 2011)*. Louvain-la-Neuve:

Presses Universitaires de Louvain.

Halliday, M. A. K. and Hasan, R. 1976. *Cohesion in English.* London: Longman.

Ishikawa, S. (ed.) 2013. *Learner corpus studies in Asia and the world - Vol.1*. Kobe: School of Language and Communication, Kobe University.

Liu, M. and Braine, G. 2005. "Cohesive features in argumentative writing produced by Chinese undergraduates". *System* 33: 623-636.

Tono, Y., Kawaguchi, Y. and Minegishi, M. (eds.) 2012. *Developmental and crosslinguistic perspectives in learner corpus research*. Amsterdam: John Benjamins.

Information structure and anaphoric links – a case study and probe

Anna Nedoluzhko
Charles University in Prague
nedoluzko@ufal.mff.cuni.cz

Eva Hajičová
Charles University in Prague
hajicova@ufal.mff.cuni.cz

> Semantics (as the study of meaning) is central to the study of communication. (p. ix) … The final category of meaning … is thematic meaning, or what is communicated by the way in which a speaker or writer organizes the message in terms of ordering, focus, and emphasis. (Leech 1974, 22).

It would be like carrying coals to Newcastle to say that Geoffrey Leech was an immemorable figure in the whole history of Corpus Linguistics. And the same is true if we try to evaluate his contribution to the study to English grammar and, in particular, to *semantics*. As the motto of our paper indicates, Leech was fully aware of the central position of semantics in the study in communication; in our contribution we devote our attention to one aspect of semantics, namely the *information structure* of the sentence, and one aspect of communication, namely the *coreferential and anaphoric links* in text. Also, following the traces of Leech's interests, we base our probe on the data of a syntactically and semantically *annotated corpus*.

Information structure of the sentence (topic-focus articulation, TFA) as developed within the functional generative description of language we subscribe to (see Sgall et al. 1986) regards the dichotomy of topic and focus as the 'aboutness' relation: the focus of the sentence says something ABOUT its topic. A sentence on its syntactico-semantic (tectogrammatical) layer is represented by a dependency tree each node of which is assigned a label identifying its underlying syntactic function (such as Actor, Patient, Addressee, etc. and different kinds of circumstantials) and a specification of *contextual boundness* or non-boundness. Based on these primary features, the sentence representation can be divided into the topic and the focus of the whole sentence. In the multilayered annotation scheme of the *Prague Dependency Treebank* (PDT) [91] TFA is captured on the underlying,

[91] PDT is an annotated collection of Czech texts, publicly available on http://ufal.mff.cuni.cz/pdt2.0, (with the data themselves available at LDC under the catalog No. LDC2006T01). It contains 3165 documents (text segments mainly of a journalistic genre) comprising of 49431 sentences

tectogrammatical layer, by a special attribute of TFA with values for contextual boundness. The PDT has served also for additional information on discourse relations and for annotation of some basic coreferential, anaphoric and associative (bridging) links (see Poláková et al. 2013, Nedoluzhko 2011).

Within the above mentioned approach to the information structure of the sentence, it is quite natural to suppose that *contextually bound* nouns and nominal groups (NG's) are linked to their antecedents by some kind of anaphoric relations, such *as grammatical coreference, textual coreference, bridging relation*, or reference to a text segment. One of the research questions we have raised then is: Is this always the case? If we find contextually bound NG's without any coreference or bridging anaphoric links, what are the reasons for this absence? Can they be reasonably classified? Are these reasons rather technical or do they have deeper theoretical background? What kind of information the existence of contextually bound NG's without anaphoric links reveals?

To answer these questions, we have used the data from the PDT briefly described in N.1. In particular, we have collected statistics of contextually bound (tectogrammatical attribute tfa="t") semantic nouns[92] that are explicitly present at the surface level (Table 1). For these nodes, we further considered how often they are linked by grammatical / textual coreference, bridging relations, or reference to a text segment. Contextually bound nodes that do not have any kind of anaphoric reference form a special class. As we can see from Table 1, almost one-third of contextually bound expressions in PDT are not linked either by coreference or by bridging relations. To find out the reasons why it is so and how does "the context" still "bind" these expressions, we excerpted 500 such cases from PDT and analyzed them from the formal, grammatical, semantic and pragmatic viewpoints. We have arrived at the following categories:

- Most frequent in our sample were cases, where contextual boundness is deduced (and marked as such, i.e. by "t") from some kind of semantic or pragmatic relation to the previous context close to bridging relations annotated in PDT, but not annotated in such a way, since an explicit specification and classification of such cases is beyond the current understanding of bridging.[93] E.g. *Ještě stále méně nákladné jsou platby za dodávky dotovaného tepla než investice do zlepšení izolačních vlastností objektů a do dalších* ***opatření-t*** *ke zlepšení tepelné pohody v nich*. (Approximate transl.: *Still less expensive are payments for the supply of energy than investments to the insolating properties of the buildings and to other* ***measures-t*** *for the improvement of temperature coziness in them*.)
- Contextual boundness of a noun group has extralinguistic reasons (e.g. expressions referring to unique objects in the given situation) or it appears as a common world-knowledge, e.g. deictic without a deictic element in ***Dvoustranu*** *připravil Jaromír Složil* (= ***The double page*** *was prepared by Jaromír Složil*)
- A noun group refers to secondary circumstances (temporal, local, etc. modifications), e.g., *U posudků* ***v minulosti*** *mohl být sebemenší náznak negativního hodnocení spouštěcím mechanismem pro šikanování.* – (Approximate transl.: ***In the past****, the slightest hint of a negative rating in the review could cause bullying*.)
- Contextually bound expressions represent rates, degrees, scales, proportions, etc., e.g., *V Indonésii je minimální denní mzda jeden a půl* ***dolaru***. – (Approximate transl.: *In Indonesia, the minimal daily salary is one and a half* ***dollar***.)
- There is also a minor group of contextually bound noun groups which are not linked by coreference or bridging relations for some other reasons, e.g. a contextually bound node is a part of an expression which is already connected by an anaphoric link, or it was technically complicated to extract the antecedent in the tectogrammatical tree, the label of contextual boundness was inserted by mistake and so on.

Not surprisingly, the distinction between these five groups of contextually bound expressions is not sharp. Moreover, the lack of anaphoric links with a contextually bound expression in PDT may be explained by more than one reason with the same instance as mentioned above.

The above analysis concerns only elements which are present at the surface level. The statistics for newly established nodes in the tectogrammatical

and 833195 occurrences of tokens (word forms and punctuation marks) annotated on all the three layers.

[92] We excerpted the following types of nominal expressions: traditional nouns and possessive adjectives (tectogrammatical attribute sempos=n.denot), deverbal nouns ending with -ní / -tí and deadjectival nouns ending with -ost (sempos=n.denot.neg), demonstrative pronouns in the positions of syntactic nouns (sempos = n.pron.def.demon) and personal pronouns and their possessive counterparts (n.pron.def.pers).

[93] For a more detailed description of bridging relations annotated in PDT see Nedoluzhko (2011).

structure (reconstructed nodes in case of ellipsis) is different. According to their definition, newly established nodes are mostly understood from the context (and, as such, should be marked as contextually bound) and are linked by anaphoric relations. Contextually non-bound new nodes are limited to list structure root nodes representing identification structures (titles) or foreign-language expressions where the TFA value is assigned to the foreign-language expression as a whole. Another group of contextually non-bound new nodes is textual ellipsis of the governing noun (*Ve své třídě české posádky obsadily první [MÍSTO] a druhé místo*. - E. transl. *Within their class the Czech teams were placed on the first [POSITION] and second positions*.). In the PDT corpus, there are 16780 contextually bound ("t") newly generated nodes; out of them there are 14464 "t" nodes with a coreferential link (NB that there may be more links from a single node), 1334 "t" nodes with a bridging relation, and 2358 nodes without any anaphoric relation . A cursory analysis of cases without anaphoric relation indicates that the reasons are similar to those for the cases occurring at the surface level. However, for newly established nodes, there are more errors in coreference annotation and also errors due to technical reasons are more frequent.

	Total in PDT	*% of total tfa=t*
contextually bound nouns	69583	100%
contextually bound nouns without any coreference or bridging links	21529	30%
contextually bound nouns with an anaphoric link:	48054	70%
contextually bound nouns with textual coreference links	37606	54%
contextually bound nouns with bridging links	6755	10%
contextually bound nouns with grammatical coreference links	3709	5%
contextually bound nouns with reference to text segment	876	1%

Table 1

In conclusion, the analysis of anaphoric relations from the point of view of their co-existence with one aspect of information structure, namely the feature of contextual boundness, has revealed that the majority of contextually bound noun groups without anaphoric links represent contextual relations of a kind different from anaphoric and basic bridging relations. Although several types of bridging relations are annotated in PDT, they cannot cover all kinds of textual cohesive interdependencies; our inquiry has pointed out one of the directions for further investigations.

Acknowledgement

We gratefully acknowledge support from the Grant Agency of the Czech Republic (grant P406/12/0658 *Coreference, discourse relations and information structure in a contrastive perspective*).

References

Bejček E., Hajičová E., Hajič J. et al. (2013): Prague Dependency Treebank 3.0. Data/software, Univerzita Karlova v Praze, MFF, ÚFAL, Prague, Czech Republic, http://ufal.mff.cuni.cz/pdt3.0/

Leech G. (1974), Semantics. Penguin Books. Harmondsworth.

Nedoluzhko A. (2011), Rozšířená textová koreference a asociační anafora, In: Studies in Computational and Theoretical Linguistics, UFAL MFF UK, Prague.

Poláková, L., Mírovský, J., Nedoluzhko, A. et al. (2013). Introducing the Prague Discourse Treebank 1.0. In: Proceedings of the 6th Int. Joint Conference onNatural Language Processing, Asian Federation of Natural Language Processing, pp. 91-99.

Sgall P., Hajičová E. and Panevová, J. (1986). The meaning of the sentence in its semantic and pragmatic aspects, ed. by J. L. Mey. Dordrecht: Reidel - Prague:Academia.

Should I say hearing-impaired or d/Deaf? A corpus analysis of divergent discourses representing the d/Deaf population in America

Lindsay C. Nickels
Lancaster University
l.nickels@lancaster.ac.uk

1 Introduction

Terminology used when referring to d/Deaf individuals in America has long been a source of strife for this community. A tight-knit, self-identified cultural and linguistic minority, this group has been characterized as defective for generations. The term 'hearing-impaired', in particular, has worn a mask of political correctness and decency despite the d/Deaf community's open repudiation of it. d/Deaf individuals, as well as advocates and allies of the d/Deaf community, believe the term 'hearing-impaired' promotes the same agenda popular years ago: one where d/Deaf people need the help of hearing people to compensate for their impairment and where the ultimate goal should be to mend said impairment in order to participate in society as a normal person would.

This clear divide in preferred terminology for the d/Deaf population in America indicates a possible divergence in discourses; however, what remains to be seen is whether or not these terms entail the use of a specific discourse that is in direct contrast to the opposing term. Therefore this analysis uses corpus techniques to explore a wide range of contemporary American English texts as a way of identifying patterns in the discourses surrounding each term. The findings will assist in determining what different discourses exist, specific aspects of said discourses, and which of them could be considered to be a discourse of hegemony.

2 Background on d/Deafness, Impairment, and Discourse

Constructions of d/Deafness have been articulated in different ways (Lane, 1995; Brueggemann, 1999; Rosen, 2003), though Rosen (2003) takes a unique approach, identifying 'jargons' used in the constructions of d/Deafness. These 'jargons' are "developed by the social institutional stakeholders that work with d/Deaf people in accordance with their agendas and practices" (p. 922) and are there to aid those social institutional stakeholders in identifying and talking about d/Deaf people. These 'jargons' stem from constellations of professions and serve to support the agenda of those institutions (Rosen, 2003, p. 923). Two of these constellations are informally referred to as the 'healing' professions (i.e. physicians, etc.) and the 'helping' professions (i.e. educators, those working in the social services, etc.); they are known formally as the 'jargons' of essentialism and social functionalism, respectively. The third constellation is made up of humanists and social scientists, those who could be considered of the critical or activist stance, and is known as agency (Rosen, 2003).

The characterizations Rosen creates in his jargons are a good match for the representations of d/Deafness in society, though the term 'jargon' does not serve much use in discourse analysis. Since these 'jargons' are reflective of social institutional constructions, which can be considered social practices, and are used both in accordance with the producer's agenda and also in identifying deaf people while they are being talked about (Rosen, 2003), it can be said that the construction is used to build the jargon, but is also represented by it. This suggests a dialectical relationship such as what is seen with discourse. As such, the three 'jargons' defined by Rosen will be considered 'discourses' in the discussion of this research.

It is in the discourse of essentialism where the normalizing paradigm so ubiquitously used to refer to the Deaf community seems to have gotten its start. This paradigm is that which encapsulates the notions of intervention and rehabilitation; and maintains the position that such a condition (i.e. d/Deafness) entails both a physical and social deficiency that prevents an individual from communicating, where the only accepted avenue for communication is an oral/aural one, and necessitates treatment to restore this individual to societal norms (Rosen, 2003).

It should be noted here that my analysis intends to use the term discourse in the spirit of Fairclough (1995), taking on the form of social action in which specific discourses are avenues for communicating and constructing social situations or positions based on the discourse producer's reality and perspective on the world. Using a discourse of essentialism/social functionalism conveys a conventionalized ideal, an ideology in which the d/Deaf population is represented as abnormal and as such unequal in the estimate of the general society. Therefore, the d/Deaf community's identity, when identified as 'hearing-impaired', appears to be situated by the hearing population, setting them apart in some way through a social representation of otherness and a discredited status in the world of normal (Oliver, 1990; Hughes, 1999; Beauchamp-Pryor, 2011). This will be demonstrated through the findings from the corpus analysis.

3 Corpus and methods

This analysis investigated the presence of discourses surrounding the representation of the Deaf community in America, specifically with the reference terms 'hearing-impaired' and 'd/Deaf', comparing and contrasting the usage of these terms in the Corpus of Contemporary American English (COCA). The COCA provides the best platform for data collection from a general corpus as it is a well-balanced representation of contemporary American English, including 450 million words between 1990-2012. Having a corpus that contains more recent texts is important to this study since the term hearing-impaired is a somewhat new term, becoming more popular in the 1990s until present after the enactment of the Americans with Disabilities Act, where this term is highlighted.

Within the COCA, I researched both terms, 'hearing-impaired' and 'd/Deaf', examining the various concordances and collocates of each. The idea behind looking at concordances and collocates in terms of discourse analysis, and in a larger sense, critical discourse analysis, was to uncover patterns within the discourse, which suggest ongoing connections between terms and perhaps with that reveal certain ideologies present in their discourse. These patterns, since they have emerged from a large range of texts and not just one individual one, may be better evidence for claiming the presence of a discourse of hegemony (Baker, 2006). The patterns discovered through these examinations shed some light on the discrepancy between each discourse and will be discussed further in the following section.

4 Findings and discussion

In looking at the results from the corpus searches, it becomes quite evident that a discourse of essentialism or social functionalism as described above appears to be employed within the concordance lines of 'hearing-impaired'. In general, the contexts in which this label is found establish a negative value judgment on the individuals being described. Specifically, they are portrayed as *disadvantaged*; are seen to be helpless; are found to be in an undesirable situation, shown with the phrase *wishing he hadn't been born hearing-impaired*; are constantly dichotomized with *normal* people; and are more often than not treated as *subjects*. Overall, the occurrences of 'hearing-impaired' within the COCA display a lack of agency and a positioning of this population that is inferior to their *normal* hearing counterparts.

Additionally 'hearing-impaired' individuals are included among groups often deemed helpless and disadvantaged in our society, those in need of some intervention or rehabilitation such as *slow learners*, the *mildly retarded*, the *learning disabled*, the *emotionally disturbed* and the *underachievers*, to name a few. There is discussion about finding a '*cure*' for their *physical challenge*. Within the list of collocates for the term 'hearing-impaired', we find *subjects* at the very top of the list, followed by *normal* and a great deal of terms associated with speech and hearing. This is not to suggest that these are the only terms found to co-occur with 'hearing-impaired', though they do suggest a pattern in discourse, one which pits this population against the whole of society and sets them apart as an 'other'.

The data collected from a search on 'd/Deaf' yielded quite different results. While there were some examples of discourses similar to what was seen with hearing-impaired, namely those where there was a focus on terms associated with speech and hearing, there were many more occurrences that presented the d/Deaf population with agency and more as a collective group. This is most notable in the collocates, which include such terms as *community*, *culture*, *language*, and *sign*. Also, the concordance lines showed 'd/Deaf' people to be dichotomized with *hearing* people, rather than *normal* people, as was seen with hearing-impaired. While the data for 'd/Deaf' did not suggest only one discourse matching that of agency, this discourse was a prominent one setting it apart from 'hearing-impaired' where this discourse was almost completely absent.

5 Concluding remarks

The above findings suggest diverging discourses in the representation of the d/Deaf population in America. It presents some convincing evidence that the term hearing-impaired can function as a discursive tool to separate this group of people from general society until they have been restored to conditions perceived as normal. While this study is not completely comprehensive in describing the exact appearance of discourses surrounding each of these labels, it does expose a clear inconsistency in how this population is represented based on which reference term is employed and therefore warrants further investigation, which I am currently undertaking. This study, as well as future research, is good testimony to show how a change in terminology can change the ideology illustrated by the discourse producer.

References

Baker, P. (2006). Using Corpora in Discourse Analysis. London: Bloomsbury Academic.

Beauchamp-Pryor, K. (2011). Impairment, cure and identity: "where do I fit in?" Disability & Society, 26(1), 5–17. doi:10.1080/09687599.2011.529662

Brueggemann, B. (1999). Lend Me Your Ear: rhetorical constructions of deafness. Washington, D.C.: Gallaudet University Press.

Fairclough, N. (1995). Critical Discourse Analysis: The Critical Study of Language. London: Longman.

Hughes, B. (1999). The Constitution of Impairment: Modernity and the aesthetic of oppression. Disability & Society, 14(2), 155–172. doi: 10.1080/09687599926244

Lane, H. (1995). Constructions of deafness. Disability & Society 10(4), 171-189.

Oliver, M. (1990). The politics of disablement: Critical texts in social work and the welfare state. London: Macmillan.

Rosen, R. S. (2003). Jargons for deafness as institutional constructions of the deaf body. Disability & Society, 18(7), 921–934. doi:10.1080/0968759032000127335

Investigating Submarine English: a pilot study

Yolanda Noguera-Díaz
Technical University of Cartagena
yolanda.noguera@upct.es

Pascual Pérez-Paredes
Universidad de Murcia
pascualf@um.es

1 Introduction

A genre is considered as a sociolinguistic activity through which members of certain discourse community achieve their communicative purposes (Swales, 1990; Bathia, 1993). Genre is then defined by its shared communicative purposes and manifested by its particular structural and linguistic features. According to Bathia (2004), a genre comprises a class of communicative events, the members of which share some set of communicative purposes.

Specialized corpora reflect the research or teaching purpose they were produced for, ad hoc purposes. They are collections of texts selected according to some common features (regarding genre and topic). Using a small corpus to obtain data for teaching a genre has also influenced language teaching to make learners aware of the relationship between the communicative purpose of a genre, the context and language chosen to achieve the purpose or as a data-driven learning (DDL).

Maritime English is the entirety of all those means of the English language which, being advisable for communication within the international maritime community, contribute to the safety of navigation and the facilitation of the seaborne business. (Trenkner 2000:77).

ME is a 'Lingua Franca' used onboard and ashore by people working in the maritime field worldwide. It entitles not only set languages – such as Seaspeak and IMO *Standard Phrases,* which are only used in ship-to-ship, ship-to-shore and shipboard communications regarding the civil world - but also, in a variety of lexical subfields, such as shipbuilding, seamanship, cargo handling, meteorology , marine engineering, electricity and electronics, automation, port operations, marine environment, safety at sea, international rules and regulations, marine insurance, shipping, business transactions, tourism and history. Even if millions of people work in this sector and need a good working knowledge of *ME* - research in this field is almost non-existent and no field-specific corpora seem to be available. The only extensive research work ever carried out was directed to produce *Seaspeak* and

IMO Standard Phrases, simplified set languages, which cover just a tiny sector of the varieties of field-specific language that can be encountered on a workplace both onboard and ashore.

2 Methodology

Submarine English can be considered a subgenre within the Maritime English scope. There are no studies explicitly addressing the variation in English used in Submarine contexts, in terms of linguistic patterns. Our aims are to establish a specialized corpus for Submarine English, which could be an important linguistic support for professional submariners in this field.

Regarding the great amount of registers and domains in which Submarine English can be divided, we will focus our attention on the Military Naval Section. One problem is that we cannot obtain submarine English texts in an easy way due to its confidential military aspect. The decision to focus only on written submarine related journal texts is primarily a pragmatic one that recognises the immediate availability of important quantities of texts at the Submarine School Library that do not require the additional layers of processing and transcription that the analysis of the spoken word requires.

Bearing in mind the difficulties to compile a submarine English corpus, we took as a starting point one year (2003) of Submarine English publications from two journals: ¨Jane's and Navy¨. We will use this pilot study to observe the possibilities of these materials and examine the lexical profile of the ten most frequent nouns. The BNC will be used to compare the first ten keywords: submarine, system, navy, class, design, boat, force, operation, missile and programme.

The Sketch Engine (Kilgarrif et al. 2004) will be our corpus tool in this pilot study. It works with all standard browsers and offers standard corpus query functions such as concordancing, sorting, filtering, etc...but it is unique in integrating grammatical making feasible to produce word sketches, that is to say, summaries of a word's grammatical and collocational behaviour. We have analyzed the above mentioned ten more frequent words in our subcorpus pilot study with the Sketch Engine, as well as the same words in the BNC.

3 First Results and conclusions.

Figure 1 gives a word sketch for the lemma noun "submarine" in both corpus giving different collocational relations such as "modifier", "modifies", "subject" and "object" grammatical relations. In our pilot corpus, its uses as a modifier, subject and object are more frequent than in the BNC. This pilot study also aims at observing a language variable like collocational frequency as a predictor of semantic specificity for the English language, as it can be determined with modifiers such as " nuclear-powered" where the statistic significance is higher in our pilot corpus 14 out of 5 in spite of the different sizes of both corpus (BNC 112.181.015 tokens versus our pilot Submarine corpus with 80.817 tokens). This a repeated effect in all ten words of study. One straight forward strategy of comparison between both corpus also shows the logdice statistics of our lemmas. Not only with the word "submarine" the logdice is higher as "subject of" in my pilot corpus (2.6 out of 2.0) but also in the rest of the studied words (logdice of ¨missile" as a subject in our corpus is 3 out of 2 in the BNC). So, these results show that our pilot study corpus of one year Submarine English research could be a milestone for future study of a decade of these journals for designing a Submarine English Corpus.

References

Aston, Guy and Burnard, Lou. The BNC handbook: exploring the British National Corpus with SARA. Edinburgh University Press, 1998.

Bloggs, J.F. and Brown, Q.V. 2004. The very complicated nature of corpus linguistics. Anytown: Anytown University Press.

Biber, Douglas, Conrad, Susan, and Reppen, Hat, H.H. 2006. *A classic research thesis with a very long title: and a subtitle*. Unpublished PhD thesis, University of Anytown.

McEnery, Tony, and Wilson, Andrew. 2001. Corpus Linguistics, 2nd ed. Edinburgh University Press

Partington, Alan. 1998. Patterns and Meanings, 1998. Amsterdam: Benjamins,

Sinclair, John.1991. Corpus, Concordance, Collocation Oxford UP.

Smith, X. 2003. "Some thoughts on submitting abstracts to conferences". In J. Jones and F.Farmer (eds.) *All about conferences*. London:

Designing English teaching activities based on popular music lyrics from a corpus perspective

Maria Claudia Nunes Delfino
São Paulo Catholic University,
São Paulo Technology College
claudia@fatecpg.com.br

1 Introduction

Popular music has been used as a tool in the teaching of foreign language for a long time (Bertoli-Dutra, 2014), but music is usually seen as an extra material to be applied when the teacher has some free time during the class or as an extracurricular activity.

In the research reported here, we argue instead that popular music can be the central element in language teaching; in fact, in our proposal, all language teaching activities were based on popular music and on texts that draw on topics related to popular music. At the same time, our goal was not to teach "pop song" English, but current spoken English. To meet this goal, the analysis of the song lyrics was used as a starting point for the materials. The patterns found in the songs as well as their register characteristics were then used as search criteria in other general English corpora and the patterns resulting from these searches were incorporated in the teaching materials as well. In short, our proposal argues for the need of a blend of register-specific and general English corpus sources. We implemented these materials in a course of English as a foreign language for elementary students enrolled at a Technology College in Brazil. In the paper presentation summarized here, we will report on the design of the corpora used in the research, present the main findings of the analysis of these corpora, and give examples of teaching activities based on the findings.

The questions which guide this work are (1) What are the high frequency lexical grammar patterns (Biber et al. 1998; Sinclair, 1991) in the pop song lyrics corpus (PSLC)*?* (2) What is the multidimensional profile (Biber, 1988) of the whole corpus as well as the individual bands in the corpus? (3) How can teaching materials be built in a way that the patterns resulted from the research be used as the starting and main point in an English class? (4) How do the students and the teacher see the process of using these corpus-based materials?

2 Metodology

The main corpus used in the analysis was a pop song lyrics corpus (PSLC), composed of around 150,000 words from 585 British and American lyrics from pop songs performed by the following bands: Beatles, Bon Jovi and Maroon 5 and the singer Bruno Mars, which was designed specifically for this project. A reference corpus was also used: COCA (ca. 450 million words), to enable the extraction of keywords. And two support corpus sources were also employed, namely corpora available on the SketchEngine portal and its recent SKELL interface.

The corpus was analyzed in two different ways. To answer the first research question, a keyword list was extracted for each musical band using COCA as the reference corpus (a wordlist from the whole COCA had been previously created in WordSmith 6 on the basis of this corpus). After that, the top 100 keywords of each band were pulled out and a concordance was run in AntConc for each keyword. Each concordance was analyzed and the major lexico-grammatical patterns were identified. The patterns sought for in the concordances were collocations, colligations, and n-grams. To answer the second research question, the corpus was tagged for over 200 different linguistic characteristics using the Biber Tagger, which is the facto morpho-syntactic tagger for MD analysis. The features were then counted with the software Biber Tag Count program, which also calculated the position of each register along each of the five major dimensions of register variation for English (Biber, 1988), which are the following: (1) Involved versus Informational Production; (2) Narrative versus Non-narrative Discourse; (3) Situation-Dependent versus Elaborated Reference; (4) Overt Expression of Argumentation and (5) Abstract versus Non-abstract Style. As a result, the relative position of each singing band along the dimensions was plotted, and their multidimensional profiles were determined. To answer the third research question, materials were designed on the basis of the research findings, and to help guide the design process, a list of "desiderata", or criteria for corpus-based material design was developed and applied (as a checklist). These criteria included items such as (1) the exercises should be based on the analysis of the relevant corpora and not on the designers' intuitions, (2) they should be replicable, (3) they should not be too time-consuming (as this would limit the production of such materials and subsequently their use in class), (4) some of their content must be fixed and some variable, (5) they should be fun, so that students feel motivated to do tasks, thereby helping increase their interest in the learning of the language, and (6) they should lead to students' empowerment, inside and outside the classroom, (7) they should be related to the students' actual or reported needs in the

workplace. A taxonomy of exercises was also developed to help in the framework for the design of the materials, which comprised: (1) blank filling, (2) sentence completion, (3) sentence unscrambling, (4) dealing with the effects of mondergreen, (5) karakoke singing, (6) concordance analysis, (7) awareness, (8) dealing with unknown words and sounds, (9) patterns relation, (10) charts analysis, among others, which resulted in the creation of several teaching units. The units incorporated additional material (concordances, frequency lists, collocate tables, etc.) found in extra corpora such as the Corpus of Contemporary American English (COCA), some SketchEngine corpora and more recently SkELL, as the need arose. To answer the fourth question, we applied the music-based activities to the students and collected their reflections on how they dealt with the issue of learning with the new materials in class; we also recorded the teacher's reflections in a reflective journal in order to gain similar insights into the process of teaching with corpus-based materials.

3 Findings

With respect to the first question, the analysis indicated that *get* and *make* (among others) were keywords. Concordances were run for these verbs and their patterns identified. As for the second research question, the analysis provided the multidimensional profile of each band. To illustrate, here is the profile for the whole corpus: Dimension 1: Very involved; Dimension 2: Not marked for narrativity; Dimension 3: Very situation dependent; Dimension 4: Very argumentative and Dimension 5: Style not abstract.

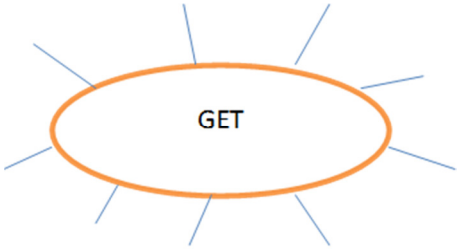
Comparing these findings with other registers afforded by the MD profiling indicated that for the whole corpus, in Dimension 1 (Involved versus Informational Production) the language used in songs tend to be between personal letters and face-to-face conversation, confirming Bertoli Dutra's (2014) findings in which she states that songs are like a conversation. The only exception in the lyrics corpus is the band Bon Jovi, whose analysis is closer to interviews and speeches and in Dimension 4 (Overt Expression of Argumentation), only this band was highly argumentative, maybe due to the high presence of modals like *should* and *could* in its lyrics which were not present in such a high incidence in the other bands.

As for the third research question, activities were designed based on the design principles guidelines and the exercise taxonomy, as mentioned. For reasons of space, these cannot be included in there, but will be reported in the presentation. Below is a short sample of the full concordance used in an activity.

1	I would gladly hit the road **get** up and go
2	Why do we let the pressure **get** into our heads?
3	**Get** down on my knees

Table 1 – Concordance Lines

As stipulated in the design guidelines, activities would ideally include visual aids. One such aid is a "word map", where students were asked to record the collocates of the word under study and then discuss / find out the meaning of the collocates or groups of collocates.



Students also explored linking sounds in pronunciation as well as some of the differences between British and American pronunciation.
Another criterion in the desiderata list was the use of activities that focused on variation. Coupled with the need for visual aids, exercises were created using the bar charts provided by COCA (shown below, for MAKE), which are quick ways in which this kind of information can be incorporated in classroom teaching (thereby fulfilling the criteria for reducing material design time).

SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
FREQ	410532	104516	73619	101129	79911	51357
PER MIL	884.13	1,093.66	814.10	1,058.29	871.27	563.95
SEE ALL SUB-SECTIONS AT ONCE						

Table 2: Chart section from COCA for the word MAKE Source: http://corpus.byu.edu/coca/

As for the fourth question, the journal provided key feedback and insights into the teaching and learning with corpus-based materials, which will be discussed in the paper presentation.

References

Berber Sardinha, T. Teaching Grammar and Corpora. In: Chapelle, C. A. (Ed.) *The Encyclopedia of Applied*

Linguistics. Blackwell Publishing Ltd., 2013

Bertoli Dutra. Multi-Dimensional Analysis of Pop Songs. In: Berber Sardinha, T. and Pinto, M. V. *Multi-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber.* John Benjamins Publishing Company. 2014

Biber, D. *Variation across Speech and Writing.* Cambridge University Press, 1988

Biber, D. Conrad, S. & Reppen, R. *Corpus Linguistics – Investigating Language Structure and Use.* Cambridge University Press, 1998.

Sinclair, J. *Corpus, Concordance, Collocation.* Oxford University Press, 1991.

Some methodological considerations when using an MD-CADS approach to track changes in social attitudes towards sexuality over time: The case of sex education manuals for British teenagers, 1950-2014

Lee Oakley
University of Birmingham
LJO848@bham.ac.uk

This paper assesses the feasibility of performing a modern diachronic corpus-assisted discourse study (henceforth MD-CADS) on a small but highly representative corpus of sex education manuals marketed at British teenagers. MD-CADS is distinguishable from other forms of corpus analysis by its emphasis on familiarity with the context of one's data, of 'reading or watching or listening to parts of the data-set, a process which can help provide a feel for how things are done linguistically in the discourse-type being studied' (Partington, Duguid & Taylor 2013: 12).

A growing number of discourse scholars have begun to address the potential of sex education materials to shape or influence young people's perceptions of sexuality. For example, Jewitt and Oyama (2001) look at the visual representation of gender and sexuality norms in sexual health posters aimed at British teenagers, whilst Chirrey (2007, 2012) investigates the advice given to young lesbians in a small set of advice manuals aimed exclusively at young lesbian women. Also, Baker (2005) discusses the construction of sexual identities in sexual health documents aimed at young gay men. Despite recent advances in the exposing of heteronormativities in high school textbooks in other countries (e.g. Wilmot and Naidoo 2014), there is still much work to be done to uncover the representations of sexuality in *sex education* advice manuals aimed at a general teenage audience in Britain.

The goal of the wider doctoral project, of which this paper is a part, is to investigate how various sexual identities are represented in British sex education manuals over a 65-year time period, starting in 1950. Much previous work has focused on the 'othering' strategies which are often used to represent gays and lesbians as strange, criminal, promiscuous, militant, shameful, etc. (Baker 2004; Baker 2014). By comparison there are relatively few diachronic studies of representations of other sexuality labels, such as bisexuality, asexuality, and most revealingly of all, heterosexuality. The research project intends to fill this gap, and in addition to this add to the growing body of linguistic work which investigates

the ways in which sexuality is represented to young people.

In order to ascertain the above, a specialized corpus of sex education texts was created (henceforth, the SexEd Corpus). The SexEd corpus was compiled by collecting sex education advice manuals which were published in Britain between 1950 and July 2014 and converting *only* the dedicated sexuality sections/chapters into plain text format. Only texts which are explicitly aimed at a general teenage readership were included (thus excluding any targeted at a particular sub-set of the teenage population, such as religious sex education manuals, young LGBT manuals, etc.)

The corpus comprises all explicit mentions of sexuality within the advice manuals. Several manuals were discarded from the analysis as they did not explicitly mention sexuality or a sexual identity label at least once. The result is a corpus of 88 texts, of 93,202 tokens, spanning the years 1950 to July 2014 (the cut-off point for data collection. See Table 1). The corpus comprises almost all of the texts of its type, and thus is a *census* rather than a representative sample. (A very small number of texts were not included due to being out of print and thus inaccessible to the researcher).

	No. of Texts	No. of Tokens
1950's	10	7,410
1960's	13	20,050
1970's	12	12,610
1980's	11	12,162
1990's	19	23,885
2000's	16	13,793
>2014	7	3,292
TOTAL	**88**	**93,202**

Table 1: Composition of the SexEd Corpus by decade

The focus of the present paper is to discuss to what extent tracking diachronic shifts in discourse prosodies is possible in a corpus where there is no recourse to comparable corpora (at present), where there is no convenient break in the data set, and where the potential for further enhancing the size and range of the corpus has been exhausted. Indeed, given that 'it is only possible to both uncover and evaluate the particular features of a discourse type by comparing it with others' (Partington *et al*. 2013: 12), this therefore presents particular difficulties when attempting to perform an MD-CADS analysis using the SexEd corpus.

Much existing (MD-)CADS work utilises ready-made corpora and/or (necessarily) arbitrary time spans in order to provide the means for comparison (e.g. Partington *et al*. 2013: 285-286, draw upon four datasets taken from the years 1993, 2005, 2009 and 2010 for their diachronic investigation of anti-Semitism tokens in the UK press). Initial attempts to 'chunk' the corpus into suitable sub-corpora based on the changing legal landscape towards non-heterosexuality would risk skewing the findings depending on which legal landmarks the researcher perceived as having significant impact on representations of sexuality. The bulk of this paper therefore addresses the range of options available to the researcher when utilising a corpus of this kind.

Acknowledgements

I would like to thank the Economic and Social Research Council (Grant number: ES/J50001X/1) for funding the wider doctoral project, of which this research is a part.

References

Baker, P. 2005. *Public Discourses of Gay Men*. London: Routledge.

Baker, P. 2014. *Using Corpora to Analyze Gender*. London: Bloomsbury.

Chirrey, D. 2007. "Women Like Us: Mediating and contesting identity in Lesbian advice literature". In, H. Sauntson and S. Kyratzis (eds.) *Language, Sexualities and Desires: Cross-Cultural Perspectives*. Basingstoke: Palgrave Macmillan, pp. 223-244.

Chirrey, D. 2012. "Reading the Script: An analysis of script formulation in coming out advice texts". *Journal of Language and Sexuality* 1 (1): 35-58.

Jewitt, C. and Oyama, R. 2001. "Visual Meaning: A Social Semiotic Approach". In, T. van Leeuwen and C. Jewitt (eds.) *Handbook of Visual Analysis*. London: Sage, pp. 134-156.

Partington, A., Duguid, A., and Taylor, C. 2013. *Patterns and Meanings in Discourse: Theory and practice in corpus-assisted discourse studies (CADS)*. Amsterdam: John Benjamins.

Wilmot, M. and Naidoo, D. 2014. "'Keeping Things Straight': The representation of sexualities in life orientation textbooks". *Sex Education* 14 (3): 32

Sharing perspectives and stance-taking in spoken learner discourse

Aisling O'Boyle
Queen's University Belfast
a.oboyle@ qub.ac.uk

Òscar Bladas
University of Barcelona
o.bladas@ alumni.ub.edu

The cooperative and 'ultra-social' nature of being human is evidenced in collective intentionality and intersubjectivity (Tomasello, 2014; 2003). In spoken discourse interactants read and interpret the intentions of others. They also signal the sharing of perspectives with one another and direct each other's inferencing processes. Such activity of sharing cognitive and social worlds takes on a new aspect when it occurs in a foreign or second language, and indeed when learning one. So, how do language learners manage or achieve the pragmatic requirements of spoken interaction? How do language learners monitor and attend to their interactants social involvement and interpretation process at the same time as regulating their own discourse production and coherence within the online constraints of face-to-face interaction? Is the demonstration of intersubjectivity a feature of fluency?

To address these questions a corpus-based study of learner discourse was undertaken. Learner corpora provide considerable opportunities to examine the nature of learner language and to investigate the variation within learner discourse and between other types of non-learner discourse. Comparisons have been made between learner or novice discourse and native speaker or expert discourse (e.g. Hyland and Milton, 1997; Gilquin et al., 2007; Gilquin and Paquot, 2007; Gilquin, 2008; Luzon, 2009; Martinez, 2005). Some see comparisons of learner and expert discourse as a means to learner empowerment (Martinez, 2005), as a means to the development of important competencies (Hyland and Milton, 1997) and as a means of overcoming non-fatal infelicities and misuse (Gilquin et al., 2007). To take up a 'difference' rather than 'deficit' model of comparison is to investigate which patterns of use occur in learner discourse, and why. As an alternative to scouring learner discourse for examples of only error or incompetency, an investigation of learner discourse can reveal something about the process of language learning itself, just as the investigation of any spontaneous speech provides clues to the process of speech production (Clark and Fox Tree, 1997; Chafe, 1994).

One way to examine how speakers share perspectives and achieve intersubjectivity is by investigating features of stance-taking. Stance-taking is described by Du Bois (2007) as:

> A public act by a social actor, achieved dialogically, through overt communicative means, of simultaneously evaluating objects, positioning subjects (self and others), and aligning with other subjects, with respect to any salient dimension of the sociocultural field (2007:163)

Markers of stance which calibrate alignment, or respond to and connect with another's perspective include pragmatic markers such as, *I think, in fact,* (Carter and McCarthy, 2006), commentary markers (Fraser, 1996), and epistemic and attitudinal lexical bundles such as, *I don't know if, I don't want to* (Biber, Conrad and Cortes, 2004).

These features of stance-taking were examined quantitatively and qualitatively in a number of spoken corpora, including: a learner corpus, a multilingual corpus and a spoken academic corpus. Search items such as stance and pragmatic markers (Carter and McCarthy, 2006) and epistemic and attitudinal lexical bundles (Biber, Conrad and Cortes, 2004) were investigated in relation to their frequency and function. In addition, examinations of whole texts where clusters of features occurred were also carried out. Comparisons were then drawn between the frequency and function of items in learner and non-learner discourse.

Findings from this study indicate that learner discourse exhibits fewer interpersonal stance markers or items which signal the learner's attention to the involvement of interactants. In addition, findings demonstrate that learner discourse displays more frequent use of textual and cognitive markers, undoubtedly indicative of the online speech constraints of the context. However, such differentials may affect the comprehension and the construction of coherence relationships in spoken learner discourse and signal a learner priority for the transaction of a message rather than a focus on engaging in the joint activity of meaning-making.

Such differences, although they may have a negative effect on interactants, are indicative of the management of spontaneous learner discourse. As learners attempt to plan, gain control, and regulate their language production together with processing the ongoing information in a face-to-face situation, elements of their message which would facilitate intention-reading or interactant interpretation may be left unattended. Or, in some instances, such interpersonal elements may be replaced by other post-hoc, non-conventional means.

The findings and discussion of this study recall the idea that "languages presuppose communication" (Harris, 1998). The implications of this study are

that it is not enough to teach interpersonal stance markers in language classrooms as 'add-ons' after language has been bifurcated into grammar and vocabulary or any other series of isolated elements for teaching purposes. It seems important to highlight that for whatever purpose and for whichever language is being learnt, that what we are attempting to do is share intentions, worlds, and minds.

References

Biber, D., Conrad, S. and Cortes, V. 2004. "'If you look at...' Lexical bundles in university teaching and textbooks". *Applied Linguistics* 25 (3):371-405.

Carter, R.A. and McCarthy, M. J. 2006. *Cambridge Grammar of English*. Cambridge University Press.

Chafe, W. 1994. *Discourse, consciousness and time.* Chicago: University of Chicago Press.

Clark, H. H. and Fox Tree, J. E. 2002. "Using uh and um in spontaneous speech". *Cognition* 84, 73-111.

Du Bois, J. 2007. "The stance triangle". In R. Englebretson (ed.) *Stancetaking in Discourse.* Amsterdam: John Benjamins.

Fraser, B. 1996. "Pragmatic markers". *Pragmatics* 6: 167-190.

Gilquin, G. 2008. "Hesitation markers among EFL learners". In J. Romero-Trillo (ed.) *Pragmatics and Corpus Linguistics: a mutalistic entente.* Berlin: Mouton de Gruyter.

Gilquin G. and Paquot M. 2007. "Spoken features in learner academic writing: identification, explanation and solution". In *Proceedings of the Fourth Corpus Linguistics Conference, University of Birmingham,* 27-30 July 2007.

Gilquin, G., Granger, S., and Paquot, M. 2007. "Learner corpora: The missing link in EAP pedagogy". *Journal of English for Academic Purposes* 6: 319–335.

Harris, R. 1998. *Introduction to integrational linguistics.* Oxford: Pergamon

Hyland, K. and Milton, J. 1997. "Qualifications and certainty in L1 and L2 students' writing". *Journal of Second Language Writing* 6: 183-205.

Luzon, M. 2009. "The use of we in a learner corpus of reports written by EFL engineering students". *Journal of English for Academic Purposes* 8 (3): 192-206.

Martinez, I.A. 2005. "Native and non-native writers' use of first person pronouns in the different sections of biology research articles in English". *Journal of Second Language Writing* 14 (3): 174-190.

Tomasello, M. 2003. *Constructing a language: a usage-based theory of language acquisition*. Harvard : Harvard University Press.

Tomasello, M. 2005. "The ultra-social animal". EuropeanJournal

Applying the concepts of Lexical Priming to German polysemantic words

Michael Pace-Sigge
University of Eastern Finland
michael.pace-sigge@uef.fi

1 Introduction

Hoey (2005) gives a detailed account of how the Theory of Lexical Priming can be seen applied to written English texts. Pace-Sigge (2013) tested the theory for spoken English and provided a background with regards to the concept of *priming* and how it was and is applied in Computational Linguistics and Psycholinguistic research (cf. Quillian: 1968).

So far, the only research into applying this theory to non-English languages seems to have been undertaken by Jantunen and Brunni (2013) as well as Hoey and Shao (forthcoming). This paper looks at one aspect of the theory only: disambiguation of polysemantic terms in German.

For this investigation, the German element of the *European Parliament Proceedings Parallel Corpus 1996-2006* as well as the collection of *German Political Speeches* of the (five) presidents of the Federal Republic 1984-2012 (Barbaresi: 2012) was used. This research focuses on the items *Steuer* and *Hut*. The words have been selected with reference to Helbig and Buscha (1984: 275). Each appears with two different genus (gender) formations. Their meanings differ accordingly: *das Steuer* refers to the "steering wheel" of a vehicle; *die Steuer*, on the other hand, means "tax". These two words are only orthographically the same. The verb-form *steuern* – "to control / to steer" - will be not looked at here.
Der Hut is "the hat"; *die Hut* refers to "care" or "guard". While the inflections change in a genus-dependent way, usage is clearly differentiated and, in particularly in the political speeches viewed, highly metaphorical. As in English, German has phrases like "that is an old hat" (*das ist ein alter Hut*) or references to people who fulfil different roles. This paper shows that there are clear differences in colligation and semantic association that go beyond the surface of genus differentiation. The Lexical Priming Theory claims that our primings will vary according to genre and domain. It is ppropriate therefore to investigate the morphological dimensions of polysemy/homonymy in a well-defined domain. Preferences for literal or metaphorical usage are inherent in the genre of political speeches investigated here. Thus, each of

the words are representing two different lexical items through a fixed framework of item-specific nestings which are both grammatical and lexical: colligational and collocational.

2 Research

Michael Hoey claims that

> (…) the patterns of one use of a polysemous word always distinguishes it from those of other uses of the word. We are (…) primed to recognise these contrastive patterns and to reproduce them. More precisely, (…) the collocations, semantic associations and colligations a word is primed for will systematically distinguish its polysemous senses. (Hoey, 2005: 81)

Such an approach to disambiguate word meanings can be traced back to early theoretical work by Quillian (1962 and 1968) who stated that "the resolution of a polysemantic ambiguity (...) consists of exploiting clues in the words, sentences or paragraphs (…) which make certain alternate meanings impossible." (Quillian 1962: 17). This, in short, describes what Hoey refers to as *nesting*. A recent, in-depth investigation into polysemy and Lexical Priming has been provided by Tsiamita (2012). With respect to morphology, however, Hoey (2005) focussed on the English language. This means that the only issue for priming discussed by him are usage patterns of singular and plural patterns (e.g consequence – consequences). Since then, Hoey has also looked at English compared to Chinese morphology. A case has been made, however to expand the "priming" theses as follows:

> (…) it seems essential to postulate a hypothesis that concerns morphological (…) priming as well. Drawing upon Hoey's (2005: 13) priming hypotheses, we postulate the following: every word is primed to occur in particular morphophonological forms; these are its paradigmatic morphological preferences.
>
> (Jantunen and Brunni, 2013: 238)

Jantunen and Brunni's extension of the theory can also be used beyond the confines of foreign language teaching: morphological priming occurs for every language that has morphological elements. Jantunen and Brunni highlight that highly inflected languages show primings that fit into a very similar pattern.

With respect to the research at hand, the investigation therefore looks how the items *der Hut; die Hut* and also *das Steuer; die Steuer* are characteristically primed for native speakers in their collocations, colligations, morphological preferences and semantic associations in order to differentiate different meaningful lexical items for different purposes in political speeches.

The issue of genus in the German language has been widely investigated (a.a. – Zubin and Köpcke 1984; Fries 2001; Bewer 2004; Marki 2008). Confusingly for learners, the definitive article does not remain stable. For example, ***die*** *Frau* (the woman) switches to *von* ***der*** *Frau* (of the woman) whereas *der* is associated with the male genus - as in *der Mann* (the man). The earliest work in German to use corpora and a method based on Quillian's work to differentiate between polysemous words is described in Zimmermann (1972).

Looking at the CHILDES (corpus of children's utterances) evidence, it can be seen that the more complex, metaphorical usage is not yet present - *Steuer* is only employed by the caregivers; *Hut* is only ever used with reference to the item of clothing.

Given that the (non-CHILDES) corpora are speeches by elected representatives, the use of *die Steuer* (tax) is strongly predominant and almost exclusively literal. The use of *das Steuer* (steering wheel) appears with only two L1 collocates - *am* and *ans* (contractions of *an dem* and *an das*) and these appear overwhelmingly with reference to alcohol or drugs.

The situation is markedly different when politicians employ the word *Hut*, which is almost always used in a transferred sense. In more than 1/3 of all cases, the figurative phrase *unter einen Hut bringen* (reconcile) is used. That would be the use for *der Hut*. *Die Hut* is used notably less frequently (around 1/6 of all uses) when the speakers say *auf der Hut sein* (to be careful / to be on guard). There are also references which can be directly translated into English, for example "wearing another hat" (*einen anderen Hut tragen);* "to raise my hat" (*ich ziehe meinen Hut*) and "old hat" (*ein alter Hut*) which are also noteworthy. Such examples highlight how key prepositions (*unter, auf*) are part of a wider, meaning-giving colligational framework.

It can be seen that the same items in German are clearly differentiated where the meaning (even the transferred meaning) differs. Not only through a different genus (which need not be the orthographic article) but also through differences in frequency and, more importantly, nesting.

3 Conclusions

Going beyond the boundaries of English language use, Lexical Priming can also be seen as relevant when looking at highly inflected languages like Finnish or German. Research undertaken indicates that polysemous words are found in environments that are not only genus-specifically divergent: they also relate a contrastive patterns of use. In other words, the nesting of each of these

indicates that the speaker uses a particular item.

This research also points to the fact that usage of rare words like *Steuer* are specific to a set of speakers like parliamentarians; furthermore, such proficient users of the language hardly refer to *Hut* in its literal sense. Instead, it occurs in transferred, metaphorical meanings. As the theory of Lexical Priming claims that speaker's primings will vary according to genre and domain, it is entirely reasonable to find morphological and metaphorical dimensions of usage which are specific to a well-defined domain.

This investigation is a first step towards deeper research to find evidence that the theory of Lexical Priming can be applied to the German language. This paper, focussing exclusively on polysemy, does highlight that further research into the issue of priming patterns in the German language is a field that can provide valuable insights.

References

Barbaresi, A. 2012. German Political Speeches. Corpus and Visualization. Second release, 03/05/12. Available online at: http://purl.org/corpus/german-speeches (last accessed 21/11/14).

Bewer, F. 2004. "Der Erwerb des Artikels als Genus-Anzeiger im deutschen Erstspracherwerb". ZAS Papers in Linguistics 33 (2): 87-140.

CHILDES Corpus, German. Available online at: http://childes.psy.cmu.edu/data/germanic/german (last accessed 21/11/14).

European Parliament Proceedings Parallel Corpus 1996-2011. German Source Release. Available online at: http://www.statmt.org/europarl/ (last accessed 21/11/14).

Fries, N. 2001. "Ist Deutsch eine schwere Sprache? Am Beispiel des Genus-Systems" Berlin: Humboldt Universität. Available online at

http://www2.rz.huberlin.de/linguistik/institut/syntax/docs/fries_ds_2000.pdf (last accessed 18/11/14)

Jantunen, J. H. and Brunni, S. 2013. "Morphology, lexical priming and second language acquisition: a corpus-study on learner Finnish". In S. Granger, G. Gilquin and F. Meunier (eds) Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Corpora and Language in Use – Proceedings 1. Louvain-la-Neuve. Presses universitaires de Louvain. 235-245.

Helbig, G. and Buscha, J. 1984. Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht.. Leibzig: VEB Verlag Enzyklopädie Leibzig.

Hoey, M. 2005. Lexical Priming. A new theory of words and language. London: Routledge.

Hoey, M. and Shao, J. (forthcoming). "'English and Chinese – two languages explained by the same theory? The odd case of a psycholinguistic theory that generates corpus-linguistic hypotheses for two unrelated languages". In Simon Smith, Bin Zou & Michael Hoey (eds) Corpus linguistics in China: theory, technology and pedagogy. Houndmills, Basingstoke: Palgrave Macmillan

Marki, M. 2008. "Zur Frage der Lehr- und Lernbarkeit des Genus der deutschen Substantive". In R. Nubert (ed.) Temeswarer Beiträge zur Germanistik. Timişoara: Mirton Verlag. 119-135.

Pace-Sigge, M. 2013. Lexical Priming in Spoken English Usage. Houndmills, Basingstoke: Palgrave Macmillan.

Scott, M. 2012, WordSmith Tools version 6, Liverpool: Lexical Analysis Software.

Quillian, R. M. 1962. "A revised design for an understanding machine". Mechanical Translation 7. 17-29.

Quillian, R. M. 1968. "Semantic Memory". In M.Minsky (ed.): Semantic Information Processing, Cambridge Mass: MIT-Press: 227-270.

Tsiamita, F. 2012. Disambiguating meaning: an examination of polysemous words within the framework of lexical priming. Unpublished PhD dissertation, University of Liverpool. Available online at
http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.569456 (last accessed 10/11/13).

Zimmermann, H.H. 1972. "Zur Konzeption der automatischen Lemmatisierung von Texten". Linguistische Arbeiten 10(12): 4-10.

Zubin, D. Köpcke, K-M. 1984. "Affect Classification in the German Gender System". Lingua 63, 41-96.

The lexical representations of metaphoricity: Understanding 'metaphoricity' through the Lexical Priming theory

Katie Patterson
University of Liverpool
k.j.patterson@liverpool.ac.uk

1 Introduction

In recent research, the term 'metaphoricity' is being increasingly adopted as a way of addressing metaphoric language from the point of view of a cline theory rather than a strict dichotomy. This paper argues that whilst a dichotomy is ineffective a term for such a complex linguistic phenomenon, the decision of whether a word or phrase is metaphoric is neither as straightforward as a single-tier cline suggests. The notion of 'metaphoric meaning' has further reaching implications on our language understanding and use than is commonly discussed. The interpretation and understanding of metaphor, like any other type of language, is highly dependent on a range of factors, both explicit and subtle. These are specific to time period, genre, environment of the speakers or writers, and context. These factors are accounted for by what Hoey (2005) terms 'lexical primings'. In addition, and on a more abstract level, personal experience and judgment are also crucial factors in addressing and understanding meaning, whether metaphoric or literal (Phillip 2011). Curiously, these are factors not often taken into consideration in current metaphoric research.

Approaching metaphor from a lexical stance, the research uses corpus methods to reveal the multi-level complexity surrounding the varieties of 'metaphoric meaning'. The intention of the study is to highlight the inadequacy of the term 'metaphor' when dealing with language behaviour. Instead, the theory of Lexical Priming (Hoey, 2005) will be adopted by way of providing an explanation for, and giving insight into, the fuzziness of 'metaphoricity'. Importantly, rather than seeing metaphoricity as something inherent within a word or phrase, this research looks instead at the idea of metaphoricity as a crack in the primings of language users, at both a collective and individual level.

2 Outline of Paper

This paper will firstly discuss some key concerns with identifying and defining metaphoricity in terms of lexical, semantic, grammatical and pragmatic manifestations. Examples of each will be extracted from a corpus. The intention is to illustrate how real-world data can benefit our stance towards metaphor identification, by exposing the fuzzy and multi-layered aspects, often hidden behind the clear and unambiguous examples drawn upon so often in research articles.

Secondly, the paper will provide an outline of the Lexical Priming theory and its role in explaining the multi-faceted aspects of metaphoricity. The theory will be illustrated with an in-depth corpus study of the items *flame* and *fire* and their various metaphoric/non-metaphoric uses.

3 Aims of the research

The aim is to address two crucial aspects of metaphoricity in particular. The first issue to be confronted is one of 'belonging': in particular, whether metaphoricity belongs to the particular word/phrase in which it is born out of, or whether instead it belong to the language user. In this latter case, the discussion focuses on the roles of creating metaphoricity (the role of the speaker/writer), and interpreting or understanding that metaphoricity (the role of the hearer/reader). Here, priming is be introduced as a means of explaining the psychological aspects of our relationship with language (metaphoric or otherwise). The discussion is illustrated with particular problematic corpus examples.

Providing evidence for this latter view (metaphoricity belonging to the language user) leads on to the second issue. This concerns the conventionality of metaphor: whether when we easily recognize a word or phrase in its metaphoric sense, the familiarity or frequency inevitably reduces the metaphoricity over time. Again, this discussion focuses on our psychological relationship with metaphors, and the extent to which we are able to recognise the metaphoricity, when not called to draw upon it for our understanding. Put simply, the more we see something, the less it draws our attention. Metaphors draw our attention through their use of exploitation or deviation from some form of linguistic norm. Over-familiarity however, would imply that we are not aware of the deviation: instead, we understand and become familiar with the metaphor as a single unit with a singular meaning, entirely independent of the literal meaning. More crucially, in our understanding of that familiar or conventional metaphor, we do not need to call upon the literal meaning to help our understanding in any way. The discussion poses the question of whether we are still aware of the metaphoricity if made to think consciously of it, or if in fact, the metaphoricity is lost from our use and understanding entirely.

Finally, drawing upon the assumption that metaphoricity belongs to the language user rather

than the words in isolation, the paper concludes that metaphoricity is a highly fluid psychologically dependent phenomenon, which has the ability to *come into* and *out of* view. It is concluded that the term is dependent upon three primary factors: the metaphor creator, the receiver, and crucially, the interaction between the two. Each of these three bound-up factors will have its own set of conditions, or primings, both on a conscious and an unconscious level.

4 The Lexical Priming Theory

Hoey's theory of Lexical Priming (2005) presents a usage-based account for both the psychological motivation behind our understanding of language and our ability to use language fluently to communicate within a given context. Presently, the theory accounts for both spoken and written language within particular domains. This research aims to present an account of how lexical priming can be extended to account for metaphoric instances of language.

In relation to the priming theory, the notion of metaphoric language as a deviation or exploitation from a linguistic norm (Hanks, 2013) is one of central importance. Further than this, the very concept of metaphor relies on the idea that words have more than one sense (Charteris-Black, 2014). It is the *haziness* of the degree to which these 'senses' of a word or phrase are lexically distinct, which lexical priming seeks to explain.

To illustrate the priming process, a writer or speaker must break some form of primings or language norms, in order to create the deviation or exploitation needed to create a metaphor. *Simultaneously*, on the part of the hearer or reader, there must be a similar process of recognising such breaks or 'cracks' (Hoey's term) in the primings, in order to understand and interpret the metaphoricity. Thus on a psychological level, it is these priming cracks within the individual, which create, transfer and maintain the metaphoricity. This theory is presented in contrast to the idea that metaphoricity is inherent and provides an explanation for the more complex behaviours involved. It is also important to consider that metaphoric language is not always created and interpreted in the same way. Thus the recognition of the distinction between the metaphor creator's role (speaker/writer) and the interpreter's role (listener/reader) is paramount to the theory.

5 The Corpus

The data is taken from a corpus of Nineteenth Century writings totalling 49 million words and primarily focuses upon the single items *flame* and *fire*. The focus is on both a quantitative and qualitative corpus analysis of the particular colligations, collocations, and semantic and pragmatic patterns associated with metaphoric and non-metaphoric instances of *flame* and *fire*. Analyses of concordance lines shows specific uses of explicit grammatical structures and patterns, such as '*old flame',* but also more implicit or abstract primings such as semantic prosody and the ability to evoke particular feelings and emotions through a projection of expected primings onto novel metaphors. This is shown with '*solitary flame'*.

6 Conclusion

Based on the findings presented, two main conclusions can be drawn. Firstly, it can be concluded that the functionality of the umbrella term 'metaphor' is often far too restrictive. Evidence is provided for this in the broad range of lexical characteristics involved in metaphoric behaviour. Secondly, the paper illustrates that metaphoricity is a dynamic process dependent on many factors. It may be characterised as an inherent quality in the language, but only once it has been put there by the speaker/writer, or the hearer/reader. This may be a conscious process or it may not be, but the important point is that the metaphoricity comes about only through the role of the language users (producers and the receivers) and their primings. Thus the research urges that concept of lexical metaphor can only be comprehensively discussed when its relationship with the language user is addressed.

In sum, the research serves to illustrate that the perspective on lexical metaphor should be re-focused on to the individual language user and both the social and psychological processes that dominate meaning and our ever-changing use of language.

References

Charteris-Black, Jonathan. (2014). *Analysing Political Speeches: Rhetoric, discourse and metaphor*. London: Palgrave-Macmillan.

Hanks, Patrick (2004). 'The Syntagmatics of Metaphor' in *International Journal of Lexicography* 17:3.

Hanks, Patrick. (2013). *Lexical Analysis*. London: John Benjamins.

Hoey, Michael (2005). *Lexical Priming*. London: Routledge.

Philip, Gill (2011). *Colouring Meaning: Collocation and Connotation in Figurative Language*. Amsterdam: John Benjamins.

Citizens and migrants: the representation of immigrants in the UK primary legislation and administration information texts (2007-2011)

Pascual Pérez-Paredes
Universidad de Murcia
pascualf@um.es

1 Introduction

According to May 2014 EUROSTAT's "Migration and migrant population statistics[94]", 1.7 million estimated immigrants arrived in the EU-27 from countries outside the EU-27 in 2012. Further 1.7 million people who resided in one of the EU Member States migrated to a different Member State that year. In 2007, the immigration rate reached 8.1 % in the UK. Six years later, The Observer[95], Sunday 13 January 2013 [3], published the following: "Over the past two decades, both immigration and emigration have increased to historically high levels, with those entering the country exceeding those leaving by more than 100,000 in every year since 1998". The headline read, "Immigration is British society's biggest problem, shows survey of public". The Observer reported that while Communities Secretary Eric Pickles prepared to disclose further efforts to aid integration, he also believed that "a mastery of English is the key to social mobility and essential if people of different generations want to get on [...] [English is] the key to uniting people and increasing their understanding of one another".

It seems that the years immediately preceding the survey carried out by the think tank British Future were a period where those taking part in the poll someway came to think that immigration was the most important cause of division, a problem, according to The Observer. The tensions reported resulted in tighter immigration controls. Vítores (2013) highlights that northern countries such as the UK place now more importance on language competence than southern countries such as Spain. For example, the Home Office set tougher language competence requirements in 2013. According to Vitores (2013), in the case of the UK, this is more a barrier policy than an integration tool. Sancho Pascual (2013:6) believes that migrations and multicultural scenarios call for a deeper understanding of the "identity markers" of those persons involved in these processes.

If the British public opinion and British immigration policy apparently became tougher and warier of immigrants around 2013, can we think that this group of people was characterized in ways that portrayed them as a "problem"? In particular, how did the UK administration represent this minority in the years preceding the British future survey?

2 Research methodology: the LADEX corpus

Language corpora have been successful in attracting the attention of researchers in discourse analysis as the computational power and flexibility of current software and web services have decisively contributed to the uptake of this methodology. Some of the social issues that have been researched using corpora include the identity of minority groups such as gay men (Baker, 2005), refugees and asylum seekers (Baker & McEnery, 2005) or muslims (Baker, Gabrielatos, & McEnerey, 2013). I have made use of the research methodology in Baker et al. (2008) and Baker, Gabrielatos, & McEnerey (2013) for the combination of Critical Discourse Analysis (CDA) and Corpus Linguistics (CL).

This research draws on the English corpus of the LADEX research initiative[96]. This corpus consists of 5 main components, totalling 10.5 million tokens. For the purposes of this paper, I have drawn on the Legislation component, 1,169,000 tokens as well as the UK Adminitration information texts component, 2,342,000 tokens. In the first subcorpus we can find all the primary legislation enacted in the realm of immigration in the 2007-2011 period, while in the second we can find informative texts produced by the UK administration for informational purposes during the same period of time. This second corpus includes most of the texts in the UK Border Agency website. The corpora were uploaded to Sketch Engine (Kilgariff et al., 2004).

3 Results and discussion

The lemma "Immigrant" was not found in the first corpus, while in the second it was only observed 7.7 times per million words, always in connection with illegal status and, curiously, the Pakistani. The lemma "migrant" is used most often in its singular form in both corpora, 829.7 times and 596.1 per million words, respectively, typically pre-modified by noun phrases that describe their status in connection with the five Visas in the context of the

[94] http://ec.europa.eu/eurostat/statistics-explained/index.php/Migration_and_migrant_population_statistics#Further_Eurostat_information

[95] http://www.theguardian.com/uk/2013/jan/13/immigration-british-society-biggest-problem

[96] http://www.theguardian.com/uk/2013/jan/13/immigration-british-society-biggest-problem

UK immigration Tier System. Migrants are for the most part portrayed as applicants of leaves to remain in the UK. In the legislation corpus migrants are mainly portrayed as fee payers and must comply with tough demands from the UK administration. However, in the information texts, they are represented by the administration as highly-skilled or high-value persons in 22% of the occurrences. The lemma "citizen" behaves in a different way. It occurs 434.5 times per million words in the legislation corpus while it is observed 1319.4 times per million words in the information texts corpus. This lemma is used most often in its singular form in both corpora, in fact it occurs very rarely in plural, and it is typically modified by an adjective that describes the origin of the person or persons involved. Contrary to migrants, citizens are almost exclusively used in connection to naturalization processes or in connection with existing regulations that can be applied in naturalization processes.

Drawing on Baker, Gabrielatos & McEnerey (2013) full word sketches were examined in order to understand the representational strategies most frequently used in both corpora. These results seem to confirm that there are aspects of the immigrants' identities that are privileged when they are referred as "migrants". These aspects tend to highlight issues connected with law enforcement and order, which may potentially play a role in streaming their identities as problem-oriented. This is not the case when they are referred as "citizens", where most collocates link this notion to that of British citizenship and the sense of community.

Acknowledgements

Research funded by FFI2011-30214 – Lenguaje de la Administración Pública en el ámbito de la extranjería: estudio multilingüe e implicaciones culturales (LADEX). Spanish Ministry of Economy and Competitiveness (Ministerio de Economía y Competitividad).

References

Baker, P. 2005. Public discourses of gay men. London: Routledge.

Baker, P. & McEnerey, T. 2005. A corpus-based approach to discourses of refugees and asylum seekers in UNI and newspaper texts. Language and Politics,4, 197-226.

Baker, P. et al. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. Discourse & Society, 19: 273-306,

Baker, P. Gabrielatos, C. & McEnerey, T. 2013. Discourse analysis and media attitudes. Cambridge: Cambridge University Press.

Fernández Vítores, D. 2013. El papel de la lengua en la configuración de la migración europea: tendencias y desencuentros. Lengua y migración, 5:2, 51-66.

Kilgarriff, A., Pavel Rychly, Pavel Smrz & David Tugwell. 2004. The Sketch Engine. Proc. Euralex. Lorient, France, July: 105-116. Reprinted in Lexicology: Critical concepts in Linguistics Hanks, editor. Routledge, 2007

Sancho Pascual, M. 2013. Dimensión lingüística de las migraciones internacionales. Lengua y migración, 5:2, 5-10.

Using Wmatrix to classify open response survey data in the social sciences: observations and recommendations

Gill Philip
University of Macerata
gill.philip@unimc.it

Lorna J. Philip
University of Aberdeen
l.philip@abdn.ac.uk

Alistair E. Philip
Chartered clinical psychologist
aephilip@waitrose.com

1 Introduction

We report here on our use of Wmatrix (Rayson 2009) and the USAS tagger (Rayson et al. 2004) as an alternative to more commonly used content analysis methods for sorting and coding open response survey data in the social sciences.

Survey-based research in the social sciences often elicits open response data which is transcribed then sorted and coded, a procedure known as content analysis (Philip and Macmillan 2005). This methodological approach may be conducted by an individual researcher or by several members of a research team who then discuss their classifications to arrive at a final, definitive coding. Two particular problems arise. Firstly it is a time-consuming method, particularly in the preferred approach when more than one researcher participates in the exercise. Secondly, it is difficult to ascertain the accuracy and consistency of the coding *within* and *between* projects, i.e. in situations where more than one set of open responses in a single questionnaire need to be coded, and where similar topics are the focus of questionnaire based data collection in a number of projects. Replication is difficult because the number of categories and the level of detail that emerge from content analysis can vary considerably from one coder to the next and from one set of responses to the next. Having a finite, fixed set of categories would therefore be helpful, as would any degree of automation of the coding procedure. It is within this context that our experimentation with Wmatrix begins.

2 Extending Wmatrix to non-linear text

The decision to try out Wmatrix in the context of coding survey data was knowingly experimental. The program is designed to give its most reliable output in running text since determining the semantic class of a word is most effectively done when it is contextualised both semantically and syntactically (Rayson et al. 2004). We were well aware that the Wmatrix output might not be useful at all, because the type of data we were interested in – open responses to survey questions – comprises discrete words and short segments of text, but we thought it worth experimenting with in any case, since any tool which can considerably reduce the number of hours spent manually coding is potentially invaluable to social sciences research. Our default option, if the Wmatrix output were to prove unsatisfactory, was to use the USAS tagger to provide us with possible codes for our data, and to manually select the most appropriate one in the context. In the event, this was only necessary in three cases – to correct wrongly-coded words, to code wrongly-spelled words, and to supply codes for uncoded words (see Section 4). This was fortunate, because deciding which of several possible codes is the best fit is a time-consuming, sometimes frustrating business – possibly more time-consuming than assigning codes from scratch (see Section 5).

3 Word frequency and conceptual centrality

Our data are responses within a word association task. They appear to be fragmentary, but the words and short phrases for each section cohere at a cognitive level (this is true in general of open-response survey data).

Word association tasks are widely used in psychology and in some areas of linguistics, and request that participants state the first thing that comes to mind when they encounter a given probe word. Typically, those words (concepts) that are most centrally related to the probe word are mentioned first, with less central words/concepts appearing lower down the list, if at all. What the researcher hopes to find in the data is that all or most respondents will supply central words/concepts, while less central words/concepts will occur with much lower frequency and with greater lexical variety. What this means in practical terms is that a semantic core should make itself strongly visible due to the constant reiteration of central words/concepts, while the full extent of the semantic dispersal of the concept – which fields it touches on, and in what proportions – is informed by the less central words/concepts. There are evident parallels here with word frequency and collocation, except that in word association the co-occurrence phenomenon of interest is more abstract, something akin to Sinclair's (1996) semantic preference. The conceptual areas can be identified on the basis of raw frequency, after the semantic tags have been applied, but it is also interesting to apply a further test, since Wmatrix makes it possible for us to do so:

a comparison of the semantic fields in our data against the semantic fields found in the BNC for the same probe word (corpus search term). This allows us to highlight the semantic areas that are significantly present in our respondents' data compared to the language in general and is of particular interest to our ongoing main study because we want to assess students' vocabulary and conceptualisations of discipline-specific key words before and after taking a degree level course in Rural Geography – an area of study where lay and professional knowledge overlap and compete. [97] Comparing open response survey data with the normative data provided by the BNC is something that – to our knowledge – no previous studies of this type have attempted. This adds a further level of robustness to our qualitative analysis of data.

4 Manual intervention

The Z category in the USAS tagset is populated with grammatical words, proper names and unrecognized words (Rayson et al 2004). This is useful since it stops them from interfering in 'proper' text analysis using semantic tagging, where the focus is on semantic areas rather than structure. Our use of Wmatrix, however, is a little different from text analysis proper, since we are working with discrete words and short text fragments. It was therefore useful for our research to re-code the Z category tags wherever possible. In particular, we needed to look closely at:

- proper names with metonymical reference (e.g. 'Range Rover' standing for off-road vehicles and the people who drive them);
- proper names with restricted (local) meaning (e.g. 'King Street', specifically King Street in Aberdeen);
- acronyms (e.g. SEPA – Scottish Environmental Protection Agency);
- dialect and regional expressions (e.g. 'doofer', synonymous with thingamajig);
- archaic or non-standard spellings (e.g. 'fayre').

After dealing with these, we were left with what we are for now calling the 'Post-Office problem'.

5 The 'Post Office' problem

Wmatrix recognizes many compound nouns and codes them as single lexical items; but it does not know all compound nouns. *Post Office* – a recurring term in our data – was one of these. It had to be manually coded from the USAS tags for *post* and *office* respectively, resulting in a final coding as Q1.2 (paper documents and writing). Ideally, the code would have been for "services", but no such code is present in the tagset. We resisted the temptation to create a new class but remain not fully convinced of the choice made since it seems overly restrictive: as well as dealing with the delivery and reception of letters and parcels, post offices are retail outlets, offer a range of financial products and provide access to official services. In rural areas the post office van, until recently, was a mode of transport which allowed people to travel between places which were not served by public transport.

At the opposite end of the scale are words which attract a mind-boggling number of codes, none of which really seem to fit. USAS finds a total of 23 possible codes for *Costa* (in our data, the coffee shop), none of which captured the 'having coffee as a social event' sense expressed by the response 'Costa with friends' [probe: SOCIAL]. Such problematic items require discussion and debate before a definitive code is agreed upon.

6 Concluding remarks

We find that Wmatrix is an extremely useful tool for the initial coding of data such as that generated by open response survey questions, due largely to its speed of processing and its overall consistency and reliability. That said, we stress that it is essential to check all the output, not only to make sure that codes have been assigned correctly, but because compounds, phrases and, in some cases, even single-word responses, may benefit from multiple coding. Recurrent miscodings (not found in our data) or Z-category dumping (as in our 'Post Office problem') can often be resolved with reference to the USAS tagset. The USAS tagger is not fail-proof, however, and the researcher(s) conducting the analysis may need to make a fresh decision on the basis of the contextual cues of the response and the probe which it relates to. However a major benefit of Wmatrix is that it highlights semantic areas that could be easily overlooked because they are not central to the object of study, e.g. 'aesthetic judgement', and this further enhances the quality of the data analysis.

References

Philip, L.J. and Macmillan, D.C. 2005. "Exploring values, context and perceptions in contingent valuation studies: the CV Market Stall technique and willingness to pay for wildlife conservation". *Journal of Environmental Planning and Management* 48 (2): 257-274.

Philip, G., Philip L.J., and Philip, A.E. 2014. "Learning as conceptual acquisition: A pilot project for measuring learning outcomes in higher education." Paper presented at AELCO-SCOLA, Badajoz (Spain), 15-18

[97] Ethical approval for the study was obtained from the University of Aberdeen College of Physical Sciences Ethical Review Committee.

October 2014.

Pragglejaz group. 2007 “MIP: a Method for identifying metaphorically used words in discourse.” *Metaphor and Symbol* 22 (1): 1-39.

Rayson, P. (2009) Wmatrix corpus analysis and comparison tool. Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix

Rayson, P., Archer, D., Piao, S. L. and McEnery, T. 2004. “The UCREL semantic analysis system.” *LREC 2004*: 7-12.

Sinclair, J.M. 1996 “The search for units of meaning.” *Textus* 9: 75-106.

Steen, G., Dorst, A., Herrmann, J., Kaal, A., Krennmayr, T. and Pasma, T. 2010. *Finding Metaphor in Grammar and Usage.* Amsterdam: John Benjamins.

Integrating corpus linguistics and GIS for the Study of Environmental Discourse

Robert Poole
University of Arizona
repoole@email.arizona.edu

This study produced a corpus-based ecological discourse analysis of press releases from Rosemont Copper Company (RCC) and blog posts from the Rosemont Mine Truth (RMT) pertaining to RCC’s proposed development of an open-pit copper mine in the Santa Rita Mountains of southern Arizona. The ecolinguistic analysis details linguistic patterns and their functions within the localized discourse of a particularly controversial environmental issue and how these grammatical and semantic features form rhetorical constellations, i.e. patterns of linguistic features performing a shared rhetorical purpose, within the interest groups’ texts. Findings show that the industry group produces a rhetoric of authority, certainty, and dominion through deployment of particular constellations of lexicogrammatical features while the contrasting linguistic elements in the environmental corpus construe uncertainty, doubt, aesthetic value, and environmental stewardship. The corpus analysis of POS and semantic tags as well as GIS mapping of place name mentions reveals that the RCC rhetoric of inevitability and certainty perpetuates and advances a corporate technocratic discourse which places humans in a role of dominion and authority over nature and the environment and is situated within international financial centers and the power these global centers confer. In contrast, the oppositional discourse forwarded by RMT projects the aesthetic value of the land and a need for responsible environmental stewardship and, as evident in the mapping of place names, is anchored within the local community and the mountains. This integration of GIS and corpus linguistics contributes to our understanding of environmental discourse and the importance of place within these debates.

The corpus for the present study consisted of two small, specialized corpora containing blog posts from an environmental organization, RMT, and press releases from the mining company, RCC. The RMT website was developed by the Tucson, Arizona-based environmental group Save the Scenic Santa Ritas (SSSR) to persuade local Arizonans to challenge the proposed copper mine. The second interest group whose texts are included in the corpus is the company, RCC, which plans to construct the mine. The corpora were analyzed through the corpus

tool Antconc (Anthony, 2014), POS and semantic tag profiles were created with WMatrix (Rayson; 2008, 2009), and a log likelihood (LL) measure was used to test significance between frequencies of POS and semantic tags. The current study applies the term *constellations* to refer to the co-occurrence of multiple lexicogrammatical features around a specific rhetorical purpose within a corpus of texts.

The LL analysis of the variation between semantic and grammatical tags of the RMT and RCC corpora revealed 72 out of approximately 200 possible tags occurred at a significantly higher frequency within the environmental texts than in the industry texts. Of these 72 items, qualitative analysis of the collocational patterns and concordance lines of these items revealed several distinct rhetorical constellations patterning through the discourse from the environmental group. The first of these constellations, and the most dominant within the data, was coded as uncertainty/doubt. In this constellation, 23 items commonly occur to produce a rhetoric of doubt and uncertainty concerning the construction of the mine, the stability of the company's finances, and the potential for harm to result from the mine's development. A second constellation consists of 13 items whose confluence produces a rhetoric of aesthetic value and environmental stewardship. The environmental texts, with much greater frequency than those of the mining company, reference living creatures and plants present in the Santa Rita Mountains while also referring to a multitude of geographical spaces such as *wetlands, ponds,* and *canyons*. This density with which geographical terms and living creatures are referenced displays an intimacy with the land and the mountains and forwards a rhetoric of stewardship.

While RMT texts were marked by frequencies of features producing a uncertainty and stewardship, the first constellation within RCC texts reflects certainty and authority. For example, the tag within the RCC texts to receive the highest LL score was *general actions/making* which includes items projecting certainty towards the building of the mine, e.g. *production, operating, pursue, construct, done, installed, implementation,* and *drilled.* A second tag also exhibits the group's authority as they repeatedly deploy the phrases *we are, we will, we have, we continue,* and *we know*. Other tags from the list are *toughness: strong/weak, comparing: similar, comparing: usual, expected, success*, as well as *future.* These tags, their representative tokens, and the concordances display RCC's authority towards the mine and what they present as its imminent construction.

A second constellation within the RCC corpus was coded for marking a discourse of dominion and economics. These items display results of the mine's construction and the economic value the mine will bring to the region. For example, the tag *education* displays the frequency of references to the positive effects the mine will bring to the area's universities, schools, and students. The construction of the mine and the environmental degradation it will cause is mitigated by the economic results the mine will engender. Additional tags identify the economic value of the mine and the many minerals that will be extracted. Not surprisingly, the list continues with tagged features noting the *jobs, careers, staff*, and *workforce* that will be a result of the mine and the technical expertise the company possesses for successful exploitation of the mountain.

Extending and enhancing the discourse analysis was the application of GIS to the study. The corpora displayed both quantitative and qualitative differences in the geographical features and place names employed by the two groups. Thus, a procedure integrating corpus linguistics and GIS in a manner similar to Gregory and Hardie (2011) was completed; place name mentions were mapped using the online mapping platform CartoDB. For RMT texts, the corpus data and maps display an emphasis on naming the protected areas, orienting the reader to the mine's location in relation to Tucson, repeated mentions of the Santa Rita Mountains, and numerous references of particular geographical features. The RMT texts and maps display an emphasis on the geographical location of the proposed mine in an effort to create a bond and relationship between the mine, the mountain, and the public. However, the place name mentions within the RCC texts are overwhelmingly the international financial centers of London, New York City, Frankfurt, Shanghai and Sydney and include far fewer mentions of local geographical places similar to those in the RMT corpus. The anchoring of the RMT texts to the local community and environment is contrasted to the positioning of the RCC texts within global financial centers and the authority and power conferred by these locations.

The corpus-based analysis reveals several rhetorical constellations conspiring in the interest groups' texts that reflect and index values, ideologies, and relationships to the mine and the mountain. For RCC, the theme pervasive in their productions is one of certainty as they project confidence to their audience of investors that although the legal process for approval required by the National Environmental Policy Act (NEPA) is yet to be resolved, approval of the mine is indeed inevitable and forthcoming. In contrast, and to perhaps balance the overt confidence of RCC, RMT produces messages that emphasize the outcome is far from decided.

The study also displays the potential for integrating corpus linguistics and GIS for the study of environmental discourse, as the maps produced display the importance of place in environmental debates and how different interest groups anchor their messages within particular geographical areas. How and where messages are anchored, the rhetorical effects of this discursive practice, and the potential for subverting these practices are important questions considered in this study.

References

Anthony, L. 2014. AntConc 3.4.3 [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Gregory, I. & Hardie, A. 2011. "Visual GISting: bringing together corpus linguistics and geographical information systems". *Literary and linguistic computing,* 26(3): 297-314.

Rayson, P. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics*. 13 (4): 519-549. DOI: 10.1075/ijcl.13.4.06ray

Rayson, P. 2009. "Wmatrix: a web-based corpus processing environment". Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/

A corpus-aided approach for the teaching and learning of rhetoric in an undergraduate composition course for L2 Writers

Robert Poole
University of Arizona
repoole@email.arizona.edu

In recent years, corpus-aided pedagogy for language teaching and learning has shifted to more functional, genre-based approaches (e.g. Charles, 2007, 2011; Henry and Roseberry, 2001; L. Flowerdew, 2003; J. Flowerdew & Wan, 2010). However, despite the continued development of top-down, discourse-sensitive corpus approaches for the classroom, it would be inaccurate to assume that corpora and corpus tools have changed the "pedagogical landscape" (Römer, 2010, p. 18) as corpus research remains "largely invisible downstream to teachers and learners" (Boulton, 2010, p. 129). Further, investigations of corpus approaches for writing instruction have been few (Ädel, 2010), and although corpus consultation for developing writing skills seems clearly beneficial, corpus study in second language (L2) writing contexts "seems not to be widely practiced", especially beyond the teaching of vocabulary and collocation (Ädel, 2010, p.40).

This slow uptake of corpus pedagogy has prompted many to rethink corpus approaches with researchers increasingly positioning corpus pedagogy as a supplement to existing methods and curriculum rather than a stand-alone pedagogical approach. Conrad, for instance, asserts that if corpora and/or corpus data are to be utilized in the classroom, corpus approaches should be integrated within existing pedagogy rather than presented in isolation (2000), serving to "complement traditional language learning resources" (Chambers (2005, p. 111). Reinhardt suggests, "(corpus) approaches are commensurable with, and can be scaffolded into, more familiar approaches that focus on learning through meaningful language use and the development of critical thinking and autonomous learning skills" (2010, p. 247).

Informed by findings in EAP/ESP pedagogical contexts and guided by the assertions of Conrad (2000), Chambers (2005), and Reinhardt (2010), this study develops and reports a corpus-aided approach for an undergraduate L2 writing classroom engaged in the study of rhetoric, a common curricular focus in undergraduate composition at U.S. universities. This study is also informed by Leech's (1997) assertion that a typical university writing assignment

could be enhanced through activities that ask students to "obtain, organize, and study real language data according to individual choice" (p. 11).

For the study, twenty-one ESL participants in an undergraduate writing class analyzed the discourse of the Rosemont Copper Mine Proposal, a controversial environmental debate regarding the development of a massive open-pit copper mine in the Santa Rita Mountains of southern Arizona. The proposed mine is fewer than 40 kilometers from our city, and the mountains can be seen by students and professors from many classrooms and open spaces on campus. This often contentious local debate began in 2006 but continues to the present and is commonly discussed in local newspapers, television broadcasts, interest group websites, and across social media. The participants closely analyzed texts and videos produced by two interest groups and completed a series of activities employing instructor-prepared corpus data. The corpus-based activities were designed to raise awareness of the patterns of keywords and rhetorical strategies within the texts connected to the local controversy while also highlighting the interconnectedness of texts within the debate.

The pedagogy was informed by the concept of convergence (Leech, 1997) and the continuum and contrastive principles (McCarthy & Carter, 1994). The convergent design facilitated classroom discussion as the participants, both individually and in groups, were able to produce similar, i.e. convergent, outcomes to the analysis of corpus data. The contrastive principle guided both the instructor's creation of the corpus and the subsequent corpus-informed activities as texts from the primary interest groups, the Rosemont Copper Company and the Save the Scenic Santa Ritas Organization, were compiled into separate sub-corpora and their output compared, analyzed, and discussed by students. Finally, while the contrastive principle guided corpus design and the resulting activities, the continuum principle informed the selection of multiple texts of varied genres for close analysis by students in the classroom. Thus, the existing departmental curriculum that emphasizes rhetorical analysis of texts was complemented and enhanced through the guided study of corpus data in a manner reflecting Leech's (1997) assertion.

The corpus-informed instructional modules included analysis of concordance lines of keywords, lists of keywords generated from the interest group corpora, and a keywords and rhetoric writing assignment. The concordance analysis and discussion required students to discuss the items which pattern within the local discourse to produce particular rhetorical effects for the respective interest groups. The students also produced their own lists of keywords they believed would occur throughout the texts, supplied the rhetorical rationale for their lists, compared their lists to the actual keywords from each group, and discussed the rhetorical motivations and strategies evident in the data. Finally, students produced keyword analysis essays that engaged with the texts and their rhetorical strategies, explained the rhetorical context that prompted and influenced their production, and provided interpretations of the corpus data and texts through a classical rhetorical framework.

Through the close study of the texts and the supplemental corpus activities, students gained insight into how arguments and rhetorical appeals are structured and patterned within texts and discourse. Complementing the program's existing curriculum, the corpus approach was designed to produce increases in rhetorical awareness amongst the first year international student writers, and the activities, as indicated in post-instruction surveys, were received positively by the participants who consistently responded to their value for enhancing their understanding of rhetoric and their ability to identify rhetorical strategies within texts. The use of local texts allowed both teacher and students to contextualize and make meaning of the corpus data while the two corpora composed of texts from opposing interest groups facilitated noticing of patterns of rhetorical strategies. The participants' study of a locally relevant issue and debate enabled the students to better understand their new city and campus community while also making advances in language awareness.

References

Ädel, A. 2010. "Using corpora to teach academic writing: Challenges for the direct approach". In M. Campoy-Cubillo, B. Bellés-Fortuño, and M. Gea-Valor (eds.) *Corpus-based approaches to English language teaching.* London: Continuum.

Boulton, A. 2010. "Learning outcomes from corpus consultation". In M. Jaén, F. Valverde, and M. Pérez (eds.) *Exploring new paths in language pedagogy.* London: Equinox.

Chambers, A. 2005. "Integrating corpus consultation in language studies". *Language learning and technology* 9 (2): 111-125.

Conrad, S. 2000. "Will corpus revolutionize grammar teaching in the 21st century?". *TESOL Quarterly* 34 (3): 548-560.

Charles, M. 2011. "Using hands-on concordancing to teach rhetorical functions: evaluation and implications for EAP writing". In A. Frankenberg-Garcia, L. Flowerdew, and G. Aston (eds.) *New trends in corpora and language learning.* London: Continuum.

Flowerdew, L. 2003. "A combined corpus and systemic-functional analysis of the problem-solution pattern in a student and professional corpus of technical writing". *TESOL Quarterly* 37 (3): 489-511.

Flowerdew, J., & Wan, A. 2010. "The linguistic and the contextual in applied genre analysis: The case of the company audit report". *ESP English for Specific Purposes* 29 (2): 78-93.

Henry, A. 2007. "Evaluating language learners' response to web-based, data-driven, genre teaching materials". *English for Specific Purposes* 26 :462-484.

Leech, G. 1997. "Teaching and language corpora: a convergence". In A. Wichmann, S. Fligelstone, T. McEnery, and G. Knowles (eds.) *Teaching and language corpora.* London: Longman.

McCarthy, M. & Carter, R. 1994. *Language as discourse*. London: Longman.

Reinhardt, J. 2010. "The potential of corpus-informed L2 pedagogy". *Studies in Hispanic and lusophone linguistics* 3 (1): 239-251.

Römer, U. 2010. "Using general and specialized corpora in English language teaching: Past, present, and future". In *Corpus=based approaches to English language teaching.* M. Campoy-Cubillo, B. Bellés-Fortuño, and M. Gea-Valor (eds.). London: Continuum.

A corpus-based discourse analytical approach to analysing frequency and impact of deviations from formulaic legal language by the ICTY

Amanda Potts
Lancaster University
a.potts@lancaster.ac.uk

1 Introduction

Legal concepts are expressed in codified language, and like all languages, this undergoes change. However, legal language has not been the basis of a long history of linguistic study like many other genres (Kjær and Palsbro 2008). This may be attributed to its perceived formulaicity; legal language does not contain as many overt examples of appraisal, stance, and subjectivity as might be found in more easily accessible text types. However (unlike in many other genres), sudden or unique deviations from 'norms' in formulaic legal language can have a range of extremely adverse effects (Kopaczyk 2013). These include: confusion regarding the nature of testimonies, inconsistencies in the application of the law, and in the most extreme cases, a lengthy and expensive appeals process. This makes it an interesting data source from both a legal and a linguistic perspective, but does not negate the fact that one might scour thousands of pages of court records to locate single instances of such unconventionalities. This is where corpus linguistic methods have something to offer.

While deviation from formulaic language is acknowledged by the court, without the use of sophisticated tools for language analysis (such as those used by corpus linguists), it is difficult to identify, quantify, and analyse exact cases. Likewise, it has proven difficult to discuss the ways in which these deviations may have impacted the Tribunal, both directly (e.g. its proceedings) and indirectly (e.g. its reputation).

In this study, I make use of corpus linguistic methods to approach legal data, acknowledging on the one hand that it is largely formulaic in style, while arguing on the other that disruptions in formulaicity are meaningful and impactful in ways that distinguishes this type and genre of data from most others.

2 Study background

Analysis is undertaken on legal discourse produced by the International Criminal Tribunal for the Former Yugoslavia (ICTY) (2009) because the discourse of this court is, in itself, unique.

Established in 1993 to investigate and prosecute war crimes committed in the Balkan region in the 1990s, the ICTY was the first war crimes court established by the United Nations, and has been the first international war crimes tribunal to have been convened since the Nuremberg/Tokyo tribunals. It has been revolutionary in some areas of international humanitarian law, having charged over 160 persons. Despite these pioneering successes, the court has been widely criticised for being ineffective and costly (Kolb 2001). Some of these inefficiencies and associated cost can be linked to deviations from formulaic language, as I indicate in this paper. Evidence is drawn from a large corpus of ICTY data.

3 Data

A variety of resources related to the ICTY are made public by the United Nations—these include documentaries, courtroom videos, and documents from both the Trials and Appeals chambers. The English versions of the Trials and Appeals have been downloaded and converted from PDF and OCR to plain text with Slavic characters preserved or re-encoded wherever possible. This resulted in a corpus exceeding 10.5 million words, drawn from 71 Trials texts and 50 Appeals texts.

4 Methods

In this study, I use three corpus linguistic methods, detailing findings and testing feasibility for wider adoption by legal practitioners. These are:

- **Frequency**: Part-of-speech-tagged wordlists have been generated and sorted by frequency, allowing me to identify the most frequent references to social actors in the corpus. Predictably, the most frequent are those dealing with court proceedings (e.g. witness and prosecution) and position within military hierarchy (police, commander). Less frequent are markers of identity indicating non-military/civilian status, and a variety denoting ethnicity, which bring into focus the nature of the conflict to varying degrees.
- **Key (contrastive) collocations**: Using SketchEngine, verbal collocations of the Tribunal itself have been analysed across the diachrony of the court's Trials and Appeals. This uncovers changing n-grams (e.g. from *becoming* a leader…to *encountering* difficulties), showing variation in negotiation of transitivity and self-conceptualisation in the texts.
- **Variations in phraseology through collocation and n-gram calculation:** Using this method, I detail common uses of *responsibility* in legal language before describing the frequency and typology of deviations in usage, e.g. *individual responsibility* (preferred, frequency: 842, logDice: 11.82) vs. *personal responsibility* (dispreferred, frequency: 23, logdice: 6.85).

Using collocation and n-gram calculation (Katz et al. 2011) to discover variations in phraseology is the most fruitful of methods but also the most complex, requiring high-level understanding of the underlying issues in the corpus. This has implications for uptake by legal professionals in future research.

5 Discussion

Corpus-based discourse analyses of legal language can offer interesting insights to both lawyers and linguists. Even the most basic corpus linguistic methods (e.g. frequency analysis) can lead to discussion on the representation of social actors in discourse. For instance, the ICTY claims that "giving victims a voice" is one of its major achievements (n.d.), and indeed, *witness* is the most frequent social actor in the corpus of Trials and Appeals. In light of criticisms of the Tribunal's efficacy, it is likewise interesting to see positive construal of agency declining throughout the diachrony represented by the documents. This was perhaps one factor contributing to the poorer perception of the ICTY in its later years. The most striking findings to-date, however, seem to be those in which variations in n-grams making up larger bundles of formulaic language lead to misunderstanding and appeal. These findings may have implications for linguists, lawyers, and even members of the United Nations committees.

Critical analyses of large bodies of legal language are relatively rare, but extremely culturally relevant. I argue that *because* of its formulaicity (and not despite it), legal language is excellent fodder for corpus-based discourse analytical inquiry.

6 Acknowledgements

This research was carried out under a Radical Futures grant awarded for collaborative work between the ESRC Centre for Corpus Approaches to Social Science (CASS) at Lancaster University and the Danish National Research Foundation's Centre for Excellence in International Courts (iCourts) at the University of Copenhagen.

References

International Tribunal for the Prosecution of Persons Responsible for Serious Violations of International Humanitarian Law Committed in the Territory of the Former Yugoslavia since 1991. *Updated Statute of the*

International Criminal Tribunal for the Former Yugoslavia (2009). United Nations Security Council. Retrieved from http://www.icty.org/x/file/Legal Library/Statute/statute_sept09_en.pdf

Katz, D. M., Nommarito, M. J., Seaman, J., Candeub, A., and Agichtein, E. (2011). Legal N-Grams? A Simple Approach to Track the 'Evolution' of Legal Language. *Proceedings of JURIX 2011: The 24th International Conference on Legal Knowledge and Information Systems*, Vienna, 2011. Available at SSRN: http://ssrn.com/abstract=1971953.

Kjær, A.L. and Palsbro, L. (2008). National Identity and Law in the Context of European Integration: The Case of Denmark. *Discourse & Society, 19*, 599-627.

Kolb, R. 2001. *The jurisprudence of the Yugoslav and Rwandan Criminal Tribunals on their jurisdiction and on international crimes*, British Yearbook of International Law: 259–315.

Kopaczyk, J. (2013). *The Legal Language of Scottish Burghs. Standarization and Lexical Bundles 1380-1560*. Oxford: Oxford University Press.

The International Criminal Tribunal for the Former Yugoslavia (n.d.). *The Tribunal's accomplishments in justice and law*. The Hague. Retrieved from http://www.icty.org/x/file/Outreach/view_from_hague/jit_accomplishments_en.pdf

Recycling and replacement as self-repair strategies in Chinese and English conversations

Lihong Quan
Guangdong University of Foreign Studies

According to Schegloff et al. (1977), repair is the treatment of recurrent problems in speaking, hearing, and understanding talk-in-interaction. There are a number of different repair types. The most frequent among them, self-initiated same-turn self-repair (henceforth self-repair), is the process by which speakers revise or repeat their prior talk at their own initiation. In recent years, this type of repair has received a considerable amount of attention, most recently also from a cross-linguistic perspective.

Prior cross-linguistic studies have shown a relationship between the typological characteristics of individual languages and patterns of self-repair. They found that some typological features, such as word order, syntactic constituent types, morphological complexity of words, etc., influence self-repair in a variety of ways (Fox et al., 1996; Rieger 2003；Bada 2010；Fox et al. 2009; Fox et al., 2010).

Prior studies found that function words are used more often than content words as the destination of recycling in English, while content words are over-represented as replaced items （Fox, Wouk, Hayashi, Fincke, et al. 2009; Bada 2010; Fox, Maschler & Uhmann 2010；Nemeth 2012）。

Fox et al.'s study (1996) demonstrated that rigidity/looseness of word order organize self-repair. In actual structural practices, English displays a rigid SVO word order, and the subject begins a tightly knit clause structure. On the contrary, Japanese word order in conversation data is rather loose. Hence, different repair strategies in the two languages arise from differences of the syntactic structures. English speakers tend to organize repair globally by recycling back to clause-initial position, while Japanese speakers usually do local repairs by only repeating or replacing the part of clause produced so far (Fox et al. 1996).

It is also found that syntactic projectability influences self-repair（Huang & Tanangkingsing 2005, Wouk, 2005）. Huang & Tanangkingsing found that the level of projectability will be reflected in differences in patterns of self-repair, in particular with more local recycling in languages where projectability is low, and more large-scale recycling in languages where projectabilty is high.

There has been only a small body of research on

self-repair in Chinese (Chui 1996；Zhang 1998；Fox et al. 2009；Yao 2010, etc.）. Chui claimed that "neither syntax nor the repair pattern conditioned repair"; rather, "quantity and lexical-form complexity" are the constraining factors. On the one hand, only the word immediately prior to the repair source tends to be recycled, regardless of its category; on the other hand, if the preceding word is in complex NP form, the recycling tends to blocked (P.367). Certainly, further studies are needed to justify this claim.

Based on the above findings, the present study attempts to examine the similarities and differences between Chinese native speakers (henceforth CNS) and English native speakers (henceforth ENS) in the practice of simple recycling and replacement. Our datasets include instances of other self-repair types, such as pre- and post-framing, additions and deletions, etc. By limiting the study to two repair types, we hope to position ourselves more effectively to understand our findings.

The rationale behind the selection of Chinese data in this study is that most prior studies have so far focussed on English, and comparative studies between CNSs and ENSs are rare. To achieve our objectives, the study aims to seek answers for the following two questions:

1) In general, what are the similarities and differences between CNSs and ENSs in their use of recycling and recycling starts?

2) What are the similarities and differences between CNSs and ENSs in their use replacement and replaced items?

For the present study, we used LOCNEC (Louvain Corpus of Native English Conversation) (Gilquin et al., 2010) as our English dataset. To enable comparison, CNCC (The Corpus of Native Chinese Conversation) was set up under the framework of LOCNEC. Under this frame, the Chinese interviewees were encouraged to speak about a set topic such as a trip, or a film/play, followed by a free discussion. In addition, the interviewees were presented with four pictures telling a story, and required to recount this. Some background information about the interviewees is provided in a special learner profile. The interviewees are all undergraduate students from the same University in China, with an average age of 21.

We coded our data for the following features: syntactic category (function or content word) of all recycling and replacement instances in the two datasets. As a result, 469& 486 recyclings and 73 & 63 replacements were extracted from the Chinese and English datasets respectively.

Regarding recycling, it is found that CNSs use much less function words as recycling starts than ENSs (51.17% for CNSs and 94.00 % for ENSs). The result of Chinese data here does not corroborate earlier predictions (Fox et al., 2010), according to which the languages with function words preceding content words show a preference for recycling back to function words rather than content words to delay the next content word due.

A further observation of the data indicate that, for CNSs, adverbs make up the highest proportion (26.23%) of all recycling starts. A striking finding here is that there is a high-frequency of verbs as recycling starts (13.86%).

In the ENS dataset, however, subject pronouns make up 53.58%, followed by determiners (17.41%). Fox et al. (2010: 2492) state that the major factors for this strong preference in English are that nearly every clause in conversation has an overt subject pronoun, and that the subject-verb complex is a deeply entrenched grammatical unit.

In terms of replacement, we can see from the data that CNSs use more content words as replaced items than ENSs (64.38% for CNSs and 33.33% for ENSs). A further observation indicates that, for CNSs, adverbs make up 24.66% of replaced items, followed by verbs (20.55%). No other category reaches this level. The ENS data shows that subject pronouns account for 28.57%, followed by determiners (15.87%). The results here justify the prior prediction (Fox et al 2010) that content words tend to be over-represented in replacements. This claim seems to hold especially true for our Chinese dataset.

To summarize, it was found that Chinese and English native speakers were similar in tending to use nouns as replaced items, and they differed in that Chinese native speakers tended to employ more content words (especially verbs and adverbs) as recycling starts and replaced items than their English counterparts did. The findings are discussed from the perspectives of looseness/rigidity of word order, morpho-syntactic structure and syntactic projectability.

The major findings of the study seem to suggest that sometimes differences across languages mask the universal patterns that organize them and sometimes language-specific features tend to exert more impact on patterns of self-repair in languages such as Chinese.

References

Bada, E. (2010). Repetitions as vocalized fillers and self-repairs in English and French interlanguages. *Journal of Pragmatics*, *42*, 1680-1688.

Chui, K. (1996). Organization of repair in Chinese conversation. *Text*, 16, 343-372.

Fox, B., et al. 2009. A cross-linguistic investigation of the site of initiation in same-turn self-repair. In Sidnell, J.

(Ed.), *Conversation Analysis: Comparative Perspectives* (pp. 60-103). Cambridge: Cambridge University Press.

Fox, B., Maschler, Y., Uhmann, S. 2010. A cross-linguistic study of self-repair: Evidence from English, German and Hebrew. *Journal of Pragmatics*, 42, 2487-2505.

Gilquin, G., De Cock, S., Granger. S. 2010. *Louvain International Database of Spoken English Interlanguage*. Belgium: Presses Universitaires de Louvain.

Huang, H., Tanangkingsing, M. 2005. Repair in verb-initial languages [J]. *Language and Linguistics,* 6(4), 575-597.

Nemeth, Z. 2012. Recycling and replacement repairs as self-initiated same-turn self-repair strategies in Hungarian. *Journal of Pragmatics,* 44, 2022-2034.

Rieger, C.L., 2003. Repetitions as self-repair strategies in English and German conversations. *Journal of Pragmatics*, 35, 47-68.

Schegloff, E., Jefferson, G., Sacks, H. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53, 361-382.

Wouk, F. 2005. The syntax of repair in Indonesian. *Discourse Studies,* 7: 237-258.

Linguistic features, L1, and assignment type: What's the relation to writing quality?

Randi Reppen
Northern Arizona University
randi.reppen@nau.edu

Shelley Staples
PurdueUniversity
staples0@purdue.edu

Within corpus-based studies of second language writing, the impact of different assignment types within academic writing (e.g., descriptive vs. argumentative essays), or the influence of the writer's first language have often been neglected (Lu, 2011). This study extends previous research by looking at linguistic features in relation to writing quality across two assignments frequently used in first year writing courses: the Rhetorical Analysis and the Extended Argumentative essay, and includes the student's first language (English, Chinese and Arabic) as a factor. Through careful linguistic analysis of many lexical and grammatical features we explore the connection of linguistic features to writing quality across assignment type and L1. The corpus used for this study consists of texts collected from a required first year university writing course at a U.S. university. This course is typical of required first year writing courses found at many U.S. universities. The classes used in this study used the same textbook and the same syllabus and have the same number and types of writing assignments. The L1 Arabic and L1 Chinese students in these classes are part of a bridge program from the university's Intensive English Program and are transitioning out of the IEP by taking this first year composition course and possibly another course in their intended discipline of study. Most of the students will enroll in disciplinary courses (i.e., not English language courses) after completing the bridge program and the course under investigation. The 400,000 word corpus balanced for even numbers of texts across the three L1s and the two assignment types.

The Biber tagger (Biber, 1988, 2006) was used to identify lexical and grammatical features used in this study. Manual tag checking ensured the accuracy of the program and changes were made using an interactive fix-tagging program developed for use in Biber and Gray (2013).. In addition, a program Tagcount (Biber, 1988, 2006) was used to count the grammatical features and norm the counts of features to ensure accurate comparability across texts and categories. A separate computer program was written to identify the most frequent lexical

items within each grammatical category as well as the number of texts in which these lexical items occurred (range).

In our preliminary investigations, Factorial ANOVAs with main, simple, and interaction effects were run to determine statistical differences across L1 background (English, Chinese, and Arabic), assignment (rhetorical analysis and argumentative essay), for the lexical and grammatical features.

The results of the factorial ANOVAs indicate that there were differences in most of the lexical and grammatical features across either L1 or assignment. First, type/token ratio was significantly different across the model and for the L1 of the writers. However, there were no significant differences across assignments and there was no interaction effect between L1 and assignment. The type/token ratio was highest for L1 English writers in both assignments, while the ratio for both Arabic L1 and Chinese L1 writers was about the same.

The phrasal features (attributive adjectives and premodifying nouns) showed the greatest effect sizes, explaining 28.1% and 23.4% of the variance in L1 and assignment respectively. Premodifying nouns were found to be significantly different across both L1 and Assignment. Figure 1 shows that premodifying nouns were used most by Chinese L1 writers and least by English L1 writers, particularly in the Extended Argument and that all three groups used more premodifying nouns in the Extended Argument than the Rhetorical Analysis. Attributive adjectives showed no difference across L1s, but were used significantly more in the Extended Argument than in the Rhetorical Analysis. There was also a significant interaction effect, which reflects the fact that L1 English writers used the most attributive adjectives in the Rhetorical Analysis while L1 Chinese writers used the most attributive adjectives in the Extended Argument. Figure 1 presents bar graphs of these two variables across L1 and task.

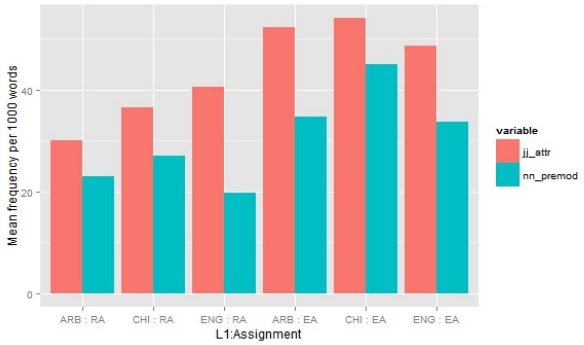


Figure 1: Premodifying adjectives and nouns across L1 and assignment

The discussion above provides some of the preliminary results of our lexical and linguistic exploration. However, we also are very interested in the relationship between lexical and linguistic features with writing quality. The development of syntactic complexity in L2 writers of English at the university level has been explored in a wide range of studies in the past years (see, e.g., Ortega, 2003 for a synthesis of 25 studies). Recently, Crossley, Salsbury, McNamara, and Jarvis (2011) investigated the relationship between both syntactic complexity and lexical measures (lexical diversity and word frequency) and essay quality for L1 writers. Both grammatical and lexical measures were significantly different between higher and lower rated essays. Taguchi, Crawford, & Wetzel (2013) also found a relationship between certain grammatical features and essay rating. Specifically, higher rated essays used more that-clause verb complements, more attributive adjectives, more post-noun-modifying prepositional phrases, and fewer subordinating conjunctions and that-relative clauses. In an effort to add to this body of research, all 240 essays in our First year writing corpus have been rated for language and organization using a holistic rubric was developed based on descriptors from the TOEFL iBT rubric (scale 0 - 5). The two areas, language and organization were rated separately and then combined for an overall score for each essay. This portion of our research, exploring the relationship of lexical and linguistic features to writing quality is currently being conducted. We expect this robust study using a carefully designed corpus and a large number of lexical and linguistic features to explore variation across L1, assignment type and writing quality to make a contribution to what is known about L1 and L2 writing.

References

Biber, D. 1988. *Variation across speech and writing*. Cambridge: CUP.

Biber, D. 2006. *University language: A corpus-based study of spoken and written registers*.Philadelphia, PA: John Benjamins Publishing.

Biber, D. and Gray, B. 2013. *Discourse characteristics of writing and speaking task types on the TOEFL*

iBT ® Test: A lexico-grammatical analysis. TOEFL iBT Research Report 19. Princeton, NJ: Educational Testing Service.

Crossley, S., Salsbury, T., McNamara, D. and Jarvis, S. 2011. Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, 28 (4) 561–580.

Lu, X. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college- level ESL writers' language development. TESOL *Quarterly*, 45, 36–42.

Ortega, L. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. Applied Linguistics 24/4: 492-518.

Taguchi, N., Crawford, W., and Wetzel, D. 2013. What Linguistic Features Are Indicative of Writing Quality? A Casse of Argumentative Essays in a College Composition Program. TESOL Quarterly, 47 (2), 420 430.

Stretching corpora to their limits: research on low-frequency phenomena

Daniel Ross
University of Illinois at Urbana-Champaign
djross3@illinois.edu

1 Introduction

Exactly what can we measure with today's corpora? Can corpora act as a proxy to specifically-designed datasets across a variety of contexts? As argued in Ross (2014), linguistic research tends to be biased toward high-frequency phenomena, meaning that we tend to only understand the most common features in languages rather than exploring the full capacity of the human language faculty. Corpus methods are one strategy to address this issue.

On the one hand, corpora are optimal for research on low-frequency phenomena because they provide direct empirical evidence in the form of millions or even billions of words. On the other hand, such large corpora are available only for a small range of linguistic varieties: usually English, usually the standard written variety, and usually contemporary usage by normal adult speakers. Sufficient corpus size is necessary for both finding relevant data and making statistical generalizations.

Therefore here I discuss the difficulties and possibilities associated with researching a particular low-frequency construction in corpora representing historical, dialectal and acquisition data for English, with implications for other languages as well. The outlook is optimistic, with such research just barely possible with the modern corpora available today.

2 The *try-and*-V construction

The *try-and*-V construction is a particular instance of a general control-verb pseudocoordination construction in English. Although several other subject control verbs such as *be sure* and *remember* can appear as the first verb in the construction (Ross 2014:211), they are too infrequent, especially in written usage, to be thoroughly investigated and statistically analyzed in most corpora. However, we can reasonably investigate the construction through its usage with *try*: the verb *try* is the 127th most frequent word in the *Corpus of Contemporary American English* (COCA: Davies 2008), with about 10 instances of *try-and*-V per million words.

Pseudocoordination has caught the attention of a number of researchers because it appears to be a mismatch between syntax (coordination) and semantics (subordination) and displays several unusual morphosyntactic properties (Ross 1967;

Culicover & Jackendoff 1997; Wiklund 2007; among others). In English, there are two types:

(1) He went and saw the movie.
(2) We will try and use corpora effectively.

The former, found with motion verbs, can be used with any morphological inflection as long as that inflection is found on both verbs (cf. Wiklund 2007). The latter, found with control verbs, may only be used in contexts with bare, uninflected verbs (Carden & Pesetsky 1977) such as imperatives, infinitives and the present-tense (except third-person-singular):

(3) We try and use corpora effectively.
(4) *He tries and use(s) corpora effectively.
(5) *We tried and use(d) corpora effectively.

This *Bare Form Condition* (BFC) can be generated by two independent properties (Ross 2013, 2014): that the second verb is a true, bare infinitive; and that the first verb must have parallel morphology to that second, necessarily uninflected verb, analogously to the requirement in motion verb pseudocoordination.

Try-and-V is frequent enough to be studied in corpora of standard English and there have been several successful studies (Lind 1983; Hommerberg & Tottie 2007; Maia 2012), which indicate that the construction is more frequent in spoken English and more frequent in British English than American English. Only in spoken British English is it used more often than *try-to*-V (about 70% of the time). Additionally, the BFC is widespread and consistent.

Below I present three case studies looking at the BFC beyond adult usage of contemporary, standard English, stretching corpora to their limits but with successful results showing that the BFC is robust.

3 Case study 1: Historical development

Although claimed to be a relatively recent phenomenon by some and dismissed as a modern error by prescriptivists, *try-and*-V has a nearly 500 year history in English having developed alongside *try-to*-V (Hommerberg & Tottie 2007; Tottie 2012).

Tottie (2012:210) claims that *try-and*-V predates *try-to*-V with raw frequencies of the sequences *try and [verb]* and *try to [verb]* in the *Early English Books Online* (EEBO) corpus, but this claim is problematic when the data is manually filtered.

The first task in research for this time period is finding a corpus with enough data; EEBO is sufficient, but without part of speech tagging this potentially ambiguous construction is challenging. The raw sequence *try and [verb]* might be normal coordination (*try and fail*), not complementation via pseudocoordination (*try and[=to] win*), with this ambiguity being the source of the construction:

(6) I will aduenture, or trie and seeke my fortune. (Baret 1573; Tottie 2012:207)

An automated search listed all instances of *try* during the 1500s in EEBO, including spelling variation, which were manually filtered to consider only those instances with *and* or *to* followed by verbs that could potentially appear to be infinitival complements. Of those, many were still ambiguous, as shown in Table 1.

try and [verb]	instances
pseudocoordination	5
ambiguous	186
not pseudocoordination	87
total	**279**
try to [verb]	instances
infinitive complement	34
ambiguous	6
not infinitive compl.	6
total	**47**

Table1: *try and/to* in EEBO 1500-1600

The results reveal that though both *try-and*-V and *try-to*-V date to the 1500s, there is no conclusive evidence that *try-and*-V is older or was more frequent at first because the majority of its instances were ambiguous during this period. We can only conclude that ambiguous contexts with *and* were more frequent than ambiguous contexts with *to*.

At this time, *try-and*-V was limited to non-finite contexts (infinitives and imperatives); the modern version of the BFC developed during the mid-1800s with present-tense usage (Ross 2013:120).

4 Case study 2: Dialectal variation

Although comparisons have been made between British English and American English, other dialects, where there might be significant variation, are more difficult to explore. The recently released *Corpus of Global Web-based English* (GloWbE: Davies 2013), with 1.9 billion words of informal written English from 20 dialects, provides an appropriate data set. After automated searching with part-of-speech tagging and manual filtering of formally ambiguous results, the BFC is shown to be ubiquitous and nearly exceptionless (Table 2).

	try-and-V	*try-to*-V
Bare	67888 (7%)	282359 (30%)
Inflected	64 (.007%)	595195 (63%)

Table2: Infinitive complements of *try* in GloWbE

Across all dialects there are only 64 instances of inflected *try* in the construction. Of these, 46 had a bare second verb, possibly by analogy to *try-to*-V. No dialect frequently uses inflected *try-and*-V. In other, smaller dialects there may still be room for variation, especially in those with non-standard

present-tense paradigms (Faarlund & Trudgill 1999) or known exceptions to the requirement for parallel inflection in motion verb pseudocoordination (Rosen 2014). Larger corpora are needed for these dialects.

5 Case study 3: Acquisition in children

The BFC is widespread and historically stable, but is it easily and consistently acquired by children? The corpora available in the CHILDES database (MacWhinney 2000) reveal that it is. No instances of inflected *try-and*-V were found in CHILDES. However, to test this statistically, a single corpus with sufficient tokens of *try-and*-V is required. Most of the corpora contained no more than two instances, but two were identified that were just large enough for this study. Both were samples of British English, where the construction is especially frequent.

First, the Fletcher corpus (Fletcher & Garman 1988; Johnson 1986) was examined, with cross-sectional data from 72 children ages 3, 5 and 7. As shown in Table 3, not only did the children not violate the BFC (statistically significant by Fisher's exact test at p<.05 for 5-7 years), but may have even acquired a categorical difference: *try-and*-V is uninflected, and *try-to*-V is inflected.

3 years	*try-and*-V	*try-to*-V
Bare	0	0
Inflected	0	4 (6)
5 years	*try-and*-V	*try-to*-V
Bare	2	0
Inflected	0	6 (10)
7 years	*try-and*-V	*try-to*-V
Bare	4 (8)	0
Inflected	0	6 (12)

Table3: *try and/to* in the Fletcher corpus
(By child, with total instances in parentheses.)

Then the Thomas corpus (Lieven, Salomo & Tomasello 2009) shows that the BFC is acquired early and consistent by a single child, recorded weekly at age 2, then monthly for ages 3 and 4. There are no violations of the BFC, and the lack of inflected *try-and*-V for ages 3 and 4, shown in Table 4, is statistically significant (p<.001).

2 years	*try-and*-V	*try-to*-V
Bare	2	0
Inflected	0	3
3 years	*try-and*-V	*try-to*-V
Bare	6	5
Inflected	0	35
4 years	*try-and*-V	*try-to*-V
Bare	15	3
Inflected	0	31

Table4: *try and/to* in the Thomas corpus

This evidence supports the *grammatical conservativity* hypothesis (Sugisaki & Snyder 2013), which states that children will make errors of omission, but few of *comission* (producing elements not found in adults speech).

6 Outlook

Research on a specific syntactic construction based on data from only a single, though frequent, verb is possible but difficult in English. But for other languages resources are needed: for example, Faroese *royna-og*-V (counterpart to *try-and*-V) may exhibit the BFC (Heycock & Petersen 2012:274), but available corpora are limited, such as *Føroyskt TekstaSavn* (about 4 million words) with only 10 instances (9 imperatives and 1 infinitive).

References

Carden, G. & Pesetsky, D. 1977. Double-Verb Constructions, Markedness, and a Fake Co-ordination. *Chicago Linguistics Society* 13: 82–82.

Culicover, P.W. & Jackendoff, R. 1997. Semantic subordination despite syntactic coordination. *Linguistic Inquiry* 28(2): 195–217.

Davies, M. 2008. The *Corpus of Contemporary American English*: 450 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.

Davies, M. 2013. *Corpus of Global Web-Based English*: 1.9 billion words from speakers in 20 countries. Available online at http://corpus2.byu.edu/glowbe/.

EEBO. Early English Books Online - Text Creation Partnership. Available online at http://www.textcreationpartnership.org/tcp-eebo/.

Faarlund, J.T. & Trudgill, P. 1999. Pseudo-coordination in English: the "try and" problem. *Zeitschrift fur Anglistik und Amerikanistik* 47(3): 210–213.

Fletcher, P. & Garman, M. 1988. Normal language development and language impairment: Syntax and beyond. *Clinical Linguistics & Phonetics* 2(2): 97–113.

Føroyskt TekstaSavn. Faroese text collection by Språkbanken and Fróðskaparsetur Føroya. Available online at http://spraakbanken.gu.se/FTS/search.phtml. (Accessed January 13th, 2015.)

Heycock, C. & Petersen, H.P. 2012. Pseudo-coordination in Faroese. In K. Braunmueller & C. Gabriel (eds.), *Multilingual Individuals and Multilingual Societies*, 259–280. Hamburg: John Benjamins.

Hommerberg, C. & Tottie, G. 2007. *Try to* or *try and*? Verb complementation in British and American English. *ICAME Journal* 31: 45–64.

Johnson, M.G. 1986. *A computer-based approach to the analysis of child language data*. Unpublished PhD thesis, University of Reading.

Lieven, E., Salomo, D. & Tomasello, M. 2009. Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics* 20(3): 481-507.

Lind, Å. 1983. The variant forms *try and*/*try to*. *English Studies* 5: 550–563.

MacWhinney, B. 2000. *The CHILDES Project: Tools for analyzing talk*. Third edition. Mahwah, NJ: Lawrence Erlbaum Associates. CHILDES available online at http://childes.psy.cmu.edu/.

Maia, J. de C. 2012. Complementation patterns of the verb *try*. *Revista Virtual dos Estudantes de Letras (ReVeLe)* 4. Available online at http://www.periodicos.letras.ufmg.br/index.php/revele/article/view/3945.

Rosen, A. 2014. *Grammatical variation and change in Jersey English*. Amsterdam: John Benjamins.

Ross, D. 2013. Dialectal variation and diachronic development of *try*-complementation. *Studies in the Linguistic Sciences: Illinois Working papers* 38: 108–147.

Ross, D. 2014. The importance of exhaustive description in measuring linguistic complexity: The case of English *try and* pseudocoordination. In F.J. Newmeyer & L.B. Preston (eds.), *Measuring Grammatical Complexity*, 202–216. Oxford: Oxford University Press.

Ross, J.R. 1967. *Constraints on Variables in Syntax*. Unpublished PhD thesis, Massachusetts Institute of Technology.

Sugisaki, K. & Snyder, W. 2013. Children's Grammatical Conservatism: New evidence. In M. Becker, J. Grinstead & J. Rothman (eds.), *Language Acquisition and Language Disorders*, 291–308. Amsterdam: John Benjamins.

Tottie, G. 2012. On the History of *try* with Verbal Complements. In S. Chevalier & T. Honegger (eds.), *Word, Words, Words: Philology and Beyond: Festschrift for Andreas Fischer on the Occasion of his 65th Birthday*, 199–214. Tübingen: Narr Francke Attempto.

Wiklund, A. 2007. *The syntax of tenselessness: tense/mood/aspect-agreeing infinitivals*. Berlin: Mouton de Gruyter.

Investigating the Great Complement Shift: a Case Study with Data from COHA

Juhani Rudanko
University of Tampere

Consider the sentences in (1a-b), both from COHA, the Corpus of Historical American English:

(1) a. Would you object to leave home?
(1890, FIC)
b. I object to signing such an order.
(1891, FIC)

In (1a) the matrix verb *object* selects a *to* infinitive complement, and in (1b) the sentential complement of the same matrix verb is what may be termed a *to -ing* complement, consisting of the preposition *to* and a following gerund. While the examples from COHA show that both types of complements were found in fairly recent English, the infinitival variant has become very rare, or even unacceptable, in current English.

The purpose of the paper is to investigate sentential complements of the matrix verb *object* during the entire time span of COHA, in order to shed light on the two types non-finite complements. To set the stage, the theoretical distinction between the two types of constructions, illustrated in (1a-b), is discussed first. Both constructions involve the word *to*, but it is argued, contrary to Duffley (2000), that only the *to* that precedes a gerund is a preposition. For its part, the *to* in *to* infinitival constructions is under the Infl node, corresponding to the Aux node in more traditional terminology. While some scholars have taken the *to* of *to* infinitives to be a semantically empty element, it is argued that this *to*, similarly to other elements under Infl, may carry a meaning.

A first objective in the empirical part of the study is to provide frequency information on the incidence of the two types of complement, as selected by the matrix verb *object*, in the last two centuries, that is, during the entire time span of COHA, up to 2009. The research tasks here are to find out how long the two complements coexisted side by side and what their frequencies were in each decade. A further task is to identify the period when the gerundial pattern came to prevail over the infinitival pattern.

A second objective is to inquire into the factors that may have played a role in favoring either type of complement during the time when both were found in reasonable numbers in the language. Questions to be investigated include the possibility of semantic differentiation of the two patterns, in the

spirit of Bolinger's Generalization, to the effect that when two constructions differ in form they may also be expected to differ in meaning (Bolinger 1968, 127). A number of concepts have been put forward as potentially having a bearing on variation between infinitival and gerundial complements in English from the point of view of meaning and conceptualization, for instance in Allerton (1988) and in later work, including Rudanko (2011), and the relevance of such factors is considered in the particular case involving the two types of complements of the matrix verb *object*.

Another factor to be investigated concerns the potential role of extractions as a determinant of alternation, taking the Extraction Principle into account, originally formulated by Vosberg (2003) and later modified in other work. One question relating to the Extraction Principle concerns the proper formulation of the Principle: it clearly pertains to the extraction of complements, as visualized in Vosberg's pioneering article, but the present study investigates the question of whether it should it also be extended to encompass the extraction of adjuncts.

The paper also relates the change in the complementation patterns of this particular verb to the broader framework of change in English that has been termed the Great Complement Shift (Rohdenburg 2006), and addresses the question of why the gerundial complement type prevailed in the case of the matrix verb *object*, and this part of the study may shed new light on the nature of the Shift.

References

Allerton, D. 1988. "Infinitivitis" in English. In Klegraf and D. Nehls, eds., *Essays on the English Language and Applied Linguistics on the Occasion of Gerhard Nickel's 60th Birthday*. Heidelberg: Julius Groos Verlag. 11-23.

Bolinger, D. 1968. Entailment and the Meaning of Structures. *Glossa* 2, 119-127.

Duffley, P. 2000. Gerund versus Infinitive as Complement of Transitive Verbs in English: the Problems of Tense and Control. *Journal of English Linguistics*, 28, 221-248.

Rohdenburg, G. 2006. The Role of Functional Constraints in the Evolution of the English Complementation System. In: C. Dalton-Puffer, D. Kastovsky, N. Ritt, and H. Schendle, eds., *Syntax, Style and Grammatical Norms: English from 1500–2000*. Bern: Peter Lang, 143–166.

Rudanko, J. 2011. *Changes in Complementation in British and American English*. Basingstoke: Palgrave Macmillan.

Vosberg, U. 2003. The Role of Extractions and *Horror Aequi* in the Evolution of *-ing* Complements with Retrospective Verbs in Modern English. In G. Rohdenburg and B. Mondorf, eds., *Determinants of Grammatical Change in English*. Berlin: Mouton de Gruyter, 305-327.

A corpus-based approach to case variation with german two-way prepositions

Jonah Rys
Ghent University
Jonah.Rys@ugent.be

1 Research Questions

The variable case marking with German two-way prepositions, which can govern both accusative (ACC) and dative (DAT) (*an, auf, in, hinter, neben, über, unter, vor, zwischen*) is traditionally assumed to coincide with an oppositional semantic difference between the expression of a directional movement (e.g. *Der Mann wandert in die$_{ACC}$ Berge* 'The man walks into the mountains') and a non-directional movement confined to a particular location (e.g. *Der Mann wandert in den$_{DAT}$ Bergen* 'The man walks around in the mountains'). However, with a particular group of "intransparent" verbs, the function of the case marking variation in their prepositional complements is much less clear-cut, e.g. aufprallen:

(1) Sie sind hart auf den$_{ACC}$ Boden der Wirklichkeit aufgeprallt.
'You have crashed violently on the rocks of reality'

(2) Das Wunderkind prallt mit fast tödlicher Wucht auf dem$_{DAT}$ Boden der Wirklichkeit auf.
'The child prodigy crashes on the rocks of reality with almost lethal force'

In recent years, several proposals have been made to account for variable case marking with these intransparent verbs (e.g. Smith 1995; Olsen 1996; Willems 2011). Most notably, Smith (1995, cf. also Duden 2006) analyzed their meanings on the basis of a source-path-goal image schema (cf. also Lakoff 1987), assuming that they inherently depict a directional, telic path. ACC marking then serves to focus on the 'path' part of the meaning (as in (1)), whilst DAT focusses on the 'endpoint' part (as in (2)).

Due to a lack of corpus-based data, however, several questions remain to be answered: 1) Can all intransparent verbs readily be analyzed as telic directional verbs with a path and an endpoint component (cf. Willems 2011, who disputes this for *verschwinden* 'to dissapear') in all or any of their senses? 2) Do individual intransparent verbs or subgroups of verbs show a preference for ACC or DAT marking, and to what extent can this preference be linked to the meaning of the verb (cf. Duden 2007; Willems 2011; Rys et al. 2014; Willems et al. to appear for indications that such preferential differences indeed exist)? 3) Are there any other linguistic factors associated with the case marking variation (e.g. diathesis, perfectivity)?

2 Methodology

In contrast to previous accounts, the present study approaches these questions from a corpus-based perspective. A corpus sample covering 30 verbs with variable case marking was set up, evenly divided over three semantically coherent groups: verbs that depict 1) a relation of inclusion (e.g. *versinken, eingraben, integrieren*), 2) attachment (e.g. *anheften, befestigen, aufhängen*), or 3) collision (e.g. *aufprallen, niederstürzen, krachen*). Sentences (300 sentences with a prepositional phrase per verb) were extracted from the German Reference Corpus (DeReKo). The search was confined to German-based newspaper texts, thus excluding Swiss and Austrian sources. This talk will primarily focus on a comparison of six collision verbs: *aufprallen*, *aufschlagen*, *auftreffen*, *auffahren*, *aufsetzen* and *krachen*.

The study consisted of three parts. First, a qualitative analysis of all sentences was conducted to delimit the senses in which every verb is used, and to verify whether these senses can indeed be described by way of conceptual spatial notions such as 'path', 'goal' and 'endpoint'. Additionally, this part of the study also served to uncover verb-specific factors that could possibly be associated to a certain case preference (e.g. the presence of a comitative adverb as in *er schlug [mit dem Kopf] auf den$_{ACC}$/dem$_{DAT}$ Boden auf* 'he hit his head on the floor').

Second, a quantitative analysis was conducted in order to check for possible semantic and morphosyntactic influences on case marking variation. All sentences were annotated for verb, case (ACC/DAT), preposition (any of the nine two-way prepositions), perfect tense (yes/no), diathesis (active, *sein*-passive, *werden*-passive), transitivity (transitive/intransitive) and the aforementioned verb senses (e.g. *aufsetzen*: 'X lands on Y' vs. 'X is based on Y') and verb-specific lexical factors. The association of these variables with case marking was evaluated by means of classification tree analysis.

Finally, a survey was conducted among native speakers to compare the results the quantitative analysis with individual acceptability judgments. To this end, 26 corpus sentences with the verbs *aufprallen*, *aufschlagen*, *auftreffen*, *auffahren* and *krachen* were presented to participants, mixed with

27 control sentences that do not allow for variable case marking. The sentences were shown as pairs, with each pair combining an ACC and a DAT marked version of the same sentence. Using a likert-type scale of 5 points, speakers were asked to compare both sentences of each pair in terms of acceptability, while allowing the 'neutral' option that both case markings were equally acceptable. The answers to the likert-type task were analyzed using chi square and fisher exact tests.

3 Results

The results of the study can be summarized in three main points.

1) A general source-path-goal image schema is not suited to describe the semantics of intransparent verbs. Particularly, the collision verbs in this study are clearly used as punctual motion verbs, if motion verbs at all, and there is no evidence that their meanings contain a 'path' and an 'endpoint' component. In other words, collision verbs clear and simple depict a punctual event of collision.

2) The quantitative analysis revealed that case marking variation with intransparent verbs is by no means completely variable. On the contrary, case marking is highly motivated by several semantic and lexical factors, most notably the verb sense (e.g. *aufprallen*: 'X crashes horizontally into an object Y': 98% ACC vs. 'X crashes downwards onto a surface Y': 87% DAT. X^2: 215,458, $p < 0,05$). Additionally, for some verbs, particular lexico-semantic factors have a large influence as well (e.g. *aufschlägen*: *Er schlug [mit dem Kopf] auf den_{ACC}/dem_{DAT} Boden auf* 'he hit his head on the floor' : 60% ACC vs. *Er schlug auf den_{ACC}/dem_{DAT} Boden auf* ,he hit the floor': 98% DAT. $X^2 = 9.5996$, $p < 0,05$). Surprisingly, there are no indications that diathesis and perfectivity, traditionally named among those factors that do influence case marking, play a significant role for any of the verbs under investigation. In any case, whatever the underlying semantic principles governing these motivations may be, it's not clear how they could be linked to a focus on the 'path' or the 'endpoint' of a trajectory

3) The results of the survey (based on 45 participants) confirm that the verb sense indeed significantly influences the acceptability of a particular case. As expected, a high frequency of a case in the corpus correlates with a high acceptance rate in the survey. However, the variation in acceptance rates between individual speakers is surprisingly large, and for the same sentences, different speakers often turned out to provide opposing ratings e.g. for

(2) Die Gesellschaft für Technische Überwachung hat festgestellt, dass der Dummy auf die_{ACC}/der_{DAT} Windschutzscheibe des Pkw aufschlug.
The Technical Inspection Organisation has determined that the dummy had hit the windscreen of the car'

29% of the respondents report a preference for ACC, as opposed to 36% of the participants preferring DAT.

Overall, the results of this case study indicate that although case marking in prepositional complements of intransparent verbs is highly motivated, the variation cannot be attributed to a single functional opposition between 'path focs' and 'endpoint focus'. On the contrary, a diverse range of different lexico-semantic factors seem to be in play, which suggests that with these verbs, the ACC-DAT opposition might only serve a 'local', verb-specific function, which differs widely on a verb-by-verb basis. The complexity and subtlety of these functions could explain why there seems to be considerable disagreement among native speakers with regards to the most acceptable case marking in a given context.

References

Duden. 2006. *Die Grammatik* (7th ed.). Mannheim: Dudenverlag.

Duden. 2007. *Richtiges und gutes Deutsch* (6th ed.). Mannheim: Dudenverlag.

Lakoff, G.. 1987. *Women, fire, and dangerous things: What categories reveal about the mind.* University of Chicago Press.

Olsen, S.. 1996. "Pleonastische Direktionale". In G. Harras and M. Bierwisch (eds.) *Wenn die Semantik arbeitet. Klaus Baumgärtner zum 65. Geburtstag*. Tübingen: Niemeyer.

Rys, J., Willems, K., De Cuypere, L.. "Akkusativ und Dativ nach Wechselpräpositionen im Deutschen. Eine Korpusanalyse von versinken, versenken, einsinken und einsenken in". In I. Doval and B. Lübke (eds.) *Raumlinguistik und Sprachkontrast. Neue Beiträge zu spatialen Relationen im Deutschen, Englischen und Spanischen*. München: Iudicium Verlag.

Smith, M. B. 1995. "Semantic motivation vs. arbitrariness in grammar: toward a more general account of the dative/accusative contrast with German two-way prepositions." In: I. Rauch and G. Carr (eds.) *Insights in Germanic linguistics I: Methodology in transition.* Berlin/New York: Mouton de Gruyter.

Willems, K. 2011. "The semantics of variable case marking (Accusative/Dative) after two-way prepositions in German locative constructions. Towards a constructionist approach". *Indogermanische Forschungen* 116: 324–366.

Willems, K. / Rys, J. / De Cuypere, L. In press. "Case alternation in argument structure constructions with

prepositional verbs. A case study in corpus-based constructional analysis". In: H. C. Boas and A. Ziem (eds.): *Constructional approaches to argument structure in German.* Berlin: Mouton de Gruyter.

Representations of the future in "accepting" and "sceptical" climate change blogs

Andrew Salway
Uni Research, Bergen
andrew.salway@uni.no

Dag Elgesem
University of Bergen
Dag.Elgesem@infomedia.uib.no

Kjersti Fløttum
University of Bergen
kjersti.flottum@if.uib.no

1 Introduction

An important part of climate change communication is how the future is conceptualised, i.e. positively or negatively, and for what and for whom (Moser and Dilling 2010), and this is an area where more research is needed (Fløttum et al. 2014). In recent years an increasing amount of communication has been taking place in the blogosphere which has been recognised as a major site for large-scale and complex discourses about climate change issues (Sharman 2014). This paper reports a corpus-assisted discourse analysis that compares representations of the future in two components of the English-language climate change blogosphere – the "accepting and "sceptical" communities.

2 Background

In previous work we initiated an investigation into how the future is represented in climate change blogs (Fløttum et al. 2014) based on the NTAP corpus (Salway et al. 2013). Corpus techniques were used to identify 18 linguistic patterns that are commonly used in future representations, e.g. 'a WORD future' (which refers to different kinds of future) and 'risk(s) | danger(s) | threat(s) to WORD' (which refers to possible future situations). By analysing instances of all these patterns it was possible to determine nine meaning categories that elucidate how people conceive of the future and the possible impacts of climate change. However, that work treated the blogosphere as homogeneous when it has been shown to contain communities with distinct viewpoints and interests, especially in the case of a complex and controversial issue like climate change (Elgesem et al. 2014). Thus, the question here is whether and how do the communities vary in the ways they conceive and represent the future.

Communities within the climate change blogosphere were classified using the NTAP corpus (Elgesem et al. 2014). Automated network analysis techniques identified groups of blogs that tend to

link to one another more than to other blogs. The 1,497 blogs from the seven largest groups at the centre of the network were classified as "accepting" the majority view on anthropogenic global warming, "sceptical" of this view, or "neutral" to it; this was done by two researchers with 84.8% inter-annotator agreement and a weighted Cohen's kappa of 0.72 which is considered to be sufficient.

Here we extend the previous analysis of future representations by comparing their presence and use in the blogs of the "accepting" and "sceptical" communities. It is thought that sceptical voices are primarily concerned with matters related to trend (questioning whether climate change is happening), attribution (questioning that human activity has an effect on climate) and/or impact (questioning that climate change has serious consequences) (Rahmstorf 2004; Dunlap and McCright 2010; Whitmarsh 2011). In other words, the "typical sceptic" thinks that the climate of the future will be pretty much like that of the past, or that changes in climate will not cause problems. So they have no reason to initiate discussion about the future: but, they may respond to such discussions in the discourses of the "accepting" community. Thus, our prediction is that future representations are more frequent in "accepting" blogs, and, when they are found in "sceptical" blogs they are being used to respond critically to the "accepting" points of view.

3 Overview of method

Table 1 describes the sub-corpora comprising "accepting" (A) and "sceptical" (S) blogs.

	Blogs	Blog posts	Words
A	775	87,311	45,467,200
S	302	35,874	28,598,738

Table 1: The "accepting" (A) and "sceptical" (S) blogs from the NTAP corpus (Salway et al. 2013).

The analysis comprised four main steps.

1. For each of the 18 patterns (Fløttum et al. 2014) the ratio of relative frequencies (RRF) (Edmundson and Wyllys 1961) was computed in order to compare the overall prevalence of future representations in the two sub-corpora. Specifically: (i) each pattern was written as a regular expression; (ii) all matches with an extra 15 characters either side were gathered and de-duplication was done, based on surrounding text, to address the problem of boilerplate in blog posts; (iii) RRF was computed as (f(A)/totalWords(A))/(f(S)/totalWords(S)); (iv) lists of blogs containing each pattern were manually scanned as an informal check for dispersion.

2. RRF values were then computed for the five most frequent fillers in each pattern for each corpus, e.g. for 'a WORD future', frequent fillers in A include 'sustainable' and 'uncertain'. This step suggested specific future representations for further investigation – those that were unusually frequent in one corpus, and those that were surprisingly similar.

3. The co-texts of selected pattern-filler combinations were subject to close reading in order to understand more about the varying perspectives and viewpoints, and differences in how they are expressed in A and S.

4. Interesting insights from 3 motivated a preliminary attempt to automatically analyse such differences in order to enable comprehensive comparisons. For a given pattern, a span of text (120 characters either side of the pattern) was gathered for all instances. The text spans from A and S were then taken as sub-corpora and RRF values for frequent words were generated. This highlighted words that are unusually frequent in the co-texts of the pattern in either A or S.

4 Main findings

From the four steps of the method.

1. All 18 patterns had an RRF > 1 meaning that all occurred relatively more frequently in A than in S which supports the prediction that A is generally more concerned about the future. However, the RRF values were spread quite evenly between 1.2-2.5 which means that some patterns are used with similar relative frequencies in A and S.

2. Stronger differences between A and S were seen for specific pattern-filler combinations, e.g. 'a sustainable future' (f(A)=231, RRF=6.1), 'a more sustainable future' (50, 15.7), 'a livable future' (35, 22.0), 'future of the Kyoto Protocol' (33, 4.2). It seems that these representations that have positive connotations for the future are not of interest to the "sceptical" community.

However, many pattern-filler combinations occur with similar relative frequencies in A and S, and a few occur at least twice as often in S, e.g. 'risk(s) | danger(s) | threat(s) to mankind', 'risk(s) | danger(s) | threat(s) to the world', 'a catastrophic future' and 'an apocalyptic future'. Our initial interpretation is that "sceptical" blogs use these representations to present the "accepting" viewpoint in value-laden terms when arguing against it.

3. The close reading of concordance lines around pattern-filler combinations such as 'threat to mankind' and 'threat to polar bears' supported this interpretation. These perceived threats are dismissed, often with overt disbelief and sarcastic tone.

4. The volume of material prevents close reading of all relevant concordance lines, so in the final step we made a preliminary attempt to capture some of these phenomena automatically. We focussed on RRF values for words in the co-texts of 'risk(s) |

danger(s) | threat(s) to WORD'. This pattern was chosen because it is one of the most frequent (f(A)=4237), f(S)=2309), and has similar relative frequency in A and S (RRF=1.2); thus it is relevant for investigating the second part of our prediction.

Words occurring relatively more around the pattern in A include 'security', 'water', 'food', 'communities' and 'ecosystems' which refer to specific concerns about the impacts of climate change. The words with high RRFs in S include 'claim', 'nonsense' and 'absurd' which suggests that bloggers are arguing against and dismissing the perceived threats of climate change, without addressing the specific concerns directly.

5 Discussion and further work

A corpus-assisted discourse analysis enabled the investigation of future representations in a large volume of material from the climate change blogosphere. Results show that, indeed, future representations are generally more prevalent in "accepting" blogs than in "sceptical" blogs. However, certain future representations are used as much, if not more, in "sceptical" blogs: in these cases it seems that they are used in order to respond critically to certain "accepting" points of view in value-laden terms. Such insights may contribute to knowledge about the human and societal dimensions of climate change, in particular about positive versus negative perspectives on the future, and thereby about what actions people are willing to engage in.

Further work will look more at the co-texts around frequent patterns that occur with similar relative frequency in A and S, in order to understand more about how and why certain future representations are used by the "sceptical" community. For this we will continue with the technique described in step 4 (above), and consider collocation analysis and framing analysis as complementary techniques. We will also add a temporal dimension to the analysis: (i) to investigate to what extent S responds to A – do we see future representations appearing in A before S?; and, (ii) to measure the change in differences between A and S over time – do we see scepticism turning into ambivalence as has been suggested elsewhere (Tvinnereim and Fløttum, submitted).

Acknowledgements

This research was supported by grants from The Research Council of Norway's SAMKUL program (LINGCLIM, project 220654) and VERDIKT program (NTAP, project 213401). We are very grateful to Knut Hofland and Lubos Steskal for their roles in creating the corpora analysed here.

References

Dunlap, R. and McCright, A. 2010. "Organized Climate Change Denial". In: R. Norgaard, D. Schlosberg, J. Dryzek eds. *The Oxford Handbook of Climate Change and Society.*

Edmundson, H.P. and Wyllys, RE. 1961. "Automatic Abstracting and Indexing - Survey and Recommendations". *Comms. of ACM* 4(5):226-234.

Elgesem, D. Steskal, L. and Diakopoulos, N. 2014. "Structure and Content of the Discourse on Climate Change in the Blogosphere: The Big Picture". *Environmental Communication*. Available online at http://www.tandfonline.com/doi/full/10.1080/17524032.2014.983536#.VJANLyusXYQ

Fløttum, K., Gjerstad, Ø., Gjesdal, A.M., Koteyko, N. and Salway, A. 2014. "Representations of the future in English language blogs on climate change". *Global Environmental Change* 29:213-222.

Moser, S. C. and Dilling, L. 2010. "Communicating Climate Change: Opportunities and Challenges for Closing the Science-Action Gap". In: R. Norgaard, D. Schlosberg, J. Dryzek eds. The Oxford Handbook of Climate Change and Society.

Rahmstorf, S. 2004. "The climate sceptics". Available online at www.pik-potsdam.de/~stefan/Publications/Other/rahmstorf_climate_sceptics_2004.pdf.

Salway, A., Touileb, S. and Hofland, K. 2013. "Applying Corpus Techniques to Climate Change Blogs". In A. Hardie and R. Love (eds.) *Corpus Linguistics 2013 Abstract Book*. Available online at http://ucrel.lancs.ac.uk/cl2013/doc/CL2013-ABSTRACT-BOOK.pdf

Sharman, A. 2014. "Mapping the climate sceptical blogosphere". *Global Environmental Change* 26:159-170.

Tvinnereim, E. and Fløttum, K. (2015). "Explaining topical prevalence in open-ended survey questions on climate change". Nature Climate Change. DOI: 10.1038/nclimate2663

Whitmarsh, L. 2011. "Scepticism and uncertainty about climate change: Dimensions, determinants and change over time", *Global Environmental Change* 21:690-700.

Developing ELT coursebooks with corpora: the case of 'Sistema Mackenzie de Ensino'

Andréa Geroldo dos Santos
Mackenzie Presbyterian University
andrea.santos@mackenzie.br

1 Introduction

This paper aims to describe how coursebooks to teach English, by Sistema Mackenzie de Ensino ('Mackenzie Educational System', hereinafter SME) – sponsored by Mackenzie Presbyterian University and Mackenzie Elementary and Secondary schools, in Brazil – have been developed based on Corpus Linguistics principles. This ELT coursebook is part of an Elementary and High school textbooks series developed and published by SME since 2011. The first edition of the whole series is to be finished until 2016, when the books for high school students will complete the series.

The option for a corpus-informed coursebook meets the needs of SME's pedagogical principles, which rely on Ausubel's Meaningful Learning theory, with roots on cognitive psychology (2000). With a corpus-informed orientation, the SME's ELT coursebooks rely on corpus observation – more specifically, on concordance lines taken from COCA [98], to introduce both lexical and grammar topics. The aim of such exercises is to help students to infer the rules and/or the usage of the topics presented as well as practise the new information studied.

The ELT series also resorts to thematic units and chapters as well as authentic texts to provide lexical and cultural input. These authentic texts, which sometimes are adapted, are analysed in one of COCA tools, called Word and Phrase.Info [99], regarding, for instance, word frequency and possible collocations.

However, such an approach is not commonly found in the ELT publishing market, let alone in the Brazilian market, in which most local produced books still tend to favour short author-made texts, word lists and rules presented, rather than inferred. Therefore, in order for the series to succeed, we felt the need to put in a teachers training with focus on presenting basic Corpus Linguistics notions, starting with a four-hour workshop on how to use COCA.

[98] Corpus of Contemporary American English, compiled by Mark Davies and available on-line at: http://corpus.byu.edu/coca/

[99] Available on-line at: http://www.wordandphrase.info/analyzeText.asp

2 Using Corpora

Part of the lexical and grammatical exercises of the SME's ELT coursebook series have been developed based on:

- DDL (Johns 1991);
- The use of concordances lines for teaching (Tribble and Jones 1997; Berber Sardinha 2004; Gavioli 2005;);
- The "three I's" - Illustration, Interaction and Induction (Carter and McCarthy, cf. Xiao and McEnery 2005), rather than using the traditional PPP approach;
- Modelling (Carter 1998).

For example, in Unit 3 – Chapter 1 of the High School book for the first grade, the topic of the chapter is healthy food. The lexical section deals with cooking verbs and contains exercises that invite students to pay attention to concordance lines from COCA and find out the collocations to each group of verbs (such as *bake cookies/pie/bread/cakes*; *boil water/potatoes/eggs*; *chop onions/vegetables/tomatoes/garlic*). The grammar section uses concordance lines from COCA (most examples from recipes) to work with the imperative.

In addition to the lexical and grammar sections, we also work with authentic texts in the reading section. These texts, linked to the topic of the units/chapters are analysed in one of COCA tools – Word and Phrase.Info, making it easier to spot the most frequent words and collocations. The tool divides the words of the texts in three different ranges, which are presented with different colours. Range 1 comprises the most common words, such as *the, with* and *take*, whereas Range 3 highlights the low frequency words, based on data from the COCA Corpus.

After this previous analysis, the words in Range 3 are included in a Glossary box next to the text, while words in Range 2 are worked in the reading comprehension exercises regarding the vocabulary understanding. To illustrate, we refer again to the case of Unit 3 – Chapter 1 of the High School book for the first grade. The analysis of the text *Food Revolution Day: exclusive interview with Jamie Oliver* provided the following:

- Words in Range 2, to be worked in comprehension exercises: *host, disease, treat* and *recipe*
- Words in Range 3, to be part of the Glossary box: *charity, supporter, life-skill, plenty, to achieve,* and *to tackle (with).*

Using corpora and corpus tools proved extremely helpful in developing exercises in the different sections of SME's ELT coursebook. Nevertheless, some issues concerning the ELT publishing market

as well as the teachers who have used/will use the books may pose a challenge for the series.

3 Challenges

Many researchers have noticed the importance of using authentic texts and corpora for developing language teaching materials (Tomlinson 2003; Mishan 2005). However, Burton (2012) points out that Corpus Linguistics has powerfully influenced the production of ELT materials, such as dictionaries and grammars, but holds less influence on the development of ELT coursebooks. That would occur because there is no motivation (or demand from the potential consumers – students and, more specifically, teachers) for publishers to innovate in such a way.

The Brazilian ELT publishing market is even more conservative, with most of the local produced coursebooks still relying on short author-made texts, list of words to be memorised, and long grammar sections with conjugation tables and mechanical exercises.

In order to face such a challenge, we believe that we should invest in teachers training (McCarthy 2008). And this is what we have done, although the series is not completed yet. The training we have devised includes the introduction of Corpus Linguistics as well as a workshop about tools that can be used for researching and teaching. For example, how to use COCA.

During those workshops, despite the teachers' interest in the training sessions, we could also observe other challenges regarding the training itself, ranging from the teachers' lack of knowledge about Corpus Linguistics, through their difficulty in dealing with computers, to their lack of time for research.

As it may be noticed, there is plenty of work to do, be it in the area of publishing or in training.

4 Concluding remarks

We believe that the partial results obtained show that, despite the challenges posed, a more effective work with corpora in order to develop ELT coursebooks is possible to be achieved. It may take time and effort, but we are in the right way.

References

Ausubel, D.P. *The Acquisition and Retention of Knowledge: A Cognitive View*. 2000. Dordrecht: Springer Science+Business Media.

Berber Sardinha, A. P. 2004. *Linguística de Corpus*. Barueri, SP: Manole.

Burton, G. 2012. "Corpora and coursebooks: destined to be strangers forever?" *Corpora* 2012 Vol. 7 (1): 91–108.

Carter, R. 1998. "Orders of reality: Cancode, communication and culture.' In: *ELT Journal*. Oxford: Oxford University Press, v.52, n.1, jan/1998, p.43-56.

Gavioli, L. 2005. *Exploring Corpora for ESP Learning*. John Benjamins Publishing. Studies in Corpus Linguistics, Vol.21.

Johns, Tim. 1991. "Should you be persuaded: two samples of data-driven learning materials". In: JOHNS, T. Johns e King, P. (eds.) Classroom Concordancing. In: *ELR Journal 4*. University of Birmingham. p.1-16.

Mishan, Freda. 2005. *Designing Authenticity into Language Learning Materials*. Bristol: Intellect Books.

McCarthy, M. 2008. *Lang. Teach*. (2008), 41:4, 563–574.

Tomlinson, B. 2003. *Developing Materials for Language Teaching*. London: Continuum.

Tribble, C. and Jones, G. 1997. *Concordances in the classroom. A resource guide for teachers*. Houston: Athelstan Publications.

XIAO, R, And Mcenery, T. 2005. *Corpora and language education*. Manuscript. Available at: http://www.corpus4u.org/archive/index.php/t-75.htm

Case in German measure Constructions

Roland Schäfer
Freie Universität Berlin
roland.schaefer@fu-berlin.de

Samuel Reichert
Freie Universität Berlin
samuel.reichert@fu-berlin.de

1. Introduction

We present a corpus study of a case alternation in German which occurs in measure constructions. These constructions contain a head noun denoting a quantity, a vessel, or a container as well as an embedded mass or count noun denoting a sort. In our corpus study, we show that the case alternation is partly influenced by easily interpretable morpho-syntactic factors. One strong semantic factor, however, is not accounted for by existing theories of measure constructions and (pseudo-)partitives.

2 German measure constructions

The basic construction is shown in (1).

(1) ein Becher Wein
a_{NOM} cup_{NOM} $wine_{NOM}$
a cup of wine

This construction is remarkable because, when the dependant sort-denoting noun forms a bare and thus indefinite NP as in (1), it strictly agrees in its case with the head noun. Semantically, these NPs are always pseudo-partitives and never partitives (cf., e.g., Barker 1998 or Vos 1999 for relevant discussion of the criteria). However, when the embedded noun comes with any determiner—be it definite or indefinite—it always has genitive case, as in (2).

(2) ein Becher des Weines
a_{NOM} cup_{NOM} the_{GEN} $wine_{GEN}$
a cup of the wine

Notice that the true partitive construction in German is formed with the preposition *von* ('of'). This *von* construction behaves much more like partitives known from English (with *of*, allowing constructions similar to English "half of the coffee", for example) compared to the genitives as in (2). The difference between (1) and (2) thus cannot be reduced to one between pseudo-partitives and partitives.
So far, we are clearly not dealing with a case alternation, because the genitive (2) and the case-agreement construction (1) are in complementary distribution. The alternation occurs only when there are embedded NPs without a determiner but with an adjunct AP, such as in (3), where (3a) and (3b) have identical meanings.

(3) a. ein Becher **leckerer** **Wein**
a_{NOM} cup_{NOM} $tasy_{NOM}$ $wine_{NOM}$
a cup of tasty wine
b. ein Becher **leckeren** **Weines**
a_{NOM} cup_{NOM} $tasy_{GEN}$ $wine_{GEN}$

Clearly, this means that the alternation occurs only when the construction is not a partitive but maximally a pseudo-partitive by virtue of not embedding a definite NP (cf. Vos 1999). Much of the literature on partitives (e.g., Anttila and Fong 2000) consequently cannot provide good clues as to what motivates the alternation. From a morpho-syntactic point of view, the odd configuration is the one illustrated in (3b), because a determinerless NP occurs in a structural position which is otherwise reserved for NPs containing a determiner as in (2). The presence of an AP generally should not license effects otherwise licensed by determiners.
Within the literature on German grammar, the construction has not received much attention, and the factors which influence the choice of genitive (3b) and case agreement (3a) have not been named. There are descriptions in reference grammars such as Eisenberg (2013), and a small study was presented in Hentschel (1993). We remedy this situation by presenting a large-scale corpus study.
In the remainder of this abstract, we refer to cases like (3a) as the "agreement construction" and cases like (3b) as the "genitive construction".

3 Design of the corpus study

In this section, we describe the data source, the sampling procedure as well as the annotation scheme.
We used the deWaC Web corpus (Baroni et al. 2009). The choice was motivated by its size (roughly 1.63 billion tokens), and by the fact that it contains texts in diverse registers, including non-standard variation. We took separate samples for embedded nouns in the three grammatical genders: 1,450 observations of masculine, 1,845 observations of neuter, and 1,719 observations of feminine embedded noun tokens. The reason for using separate samples is that German nouns show many case syncretisms, to the effect that only the masculine singular NP of the form [AP N] still differentiates between the four cases of German. In the neuter and the feminine singular, nominative and accusative are conflated. In the feminine singular, dative and genitive are conflated as well, effectively making the feminine system a two case system. We therefore focus here on the results for masculine and neuter nouns for reasons of greater clarity, although we did also look at feminine nouns, and the results pointed

into the same direction.

We calculated binomial Generalized Linear Models (GLM, e.g., Fahrmeir et al. 2013), because we were interested in several factors which might determine the choice of the agreement vs. the genitive construction—clearly a binomial response variable. As regressors, we manually annotated all 5,014 observations for the case of the quantifying noun (QNCASE=[NOM; ACC; DAT]).[100] We annotated all observations for whether the quantifying noun was definite or not (QNDEF=[1;0]), and whether it was graphemically abbreviated as in *kg* for *kilogram* (QNABBR=[1;0]). To see whether the frequency of the embedded noun plays a role, we extracted the log-frequency per one million tokens from deWaC and used it as an interval scale regressor (LOGFREQ).

4 Results and discussion

The results of the GLMs are summarized in Table 1 and Table 2.

Regressor	Odds ratio	p
(Intercept)	0.14	< 0.001 ***
QNCASE=NOM	1.37	< 0.1 —
QNCASE=DAT	1.37	< 0.05 *
QNDEF=1	4.36	< 0,001 ***
LOGFREQ	1.03	< 0.001 ***
QNABBR=1	0.05	< 0.001 ***

Table 1: GLM results for masculine sort nouns

In both models, the intercept models the most frequent level of the factors (QNCASE=ACC, QNDEF=0, QNABBR=0) as well as LOGFREQ=0. Odds ratios above 1 indicate a preference for the genitive and odds ratios below 1 indicate a preference for the agreement construction. The fact that the intercept has a very low odds ratio is indicative of the fact that the genitive construction is much rarer that the agreement construction. The genitive construction occurs with only 23.38% of the observations with masculine and 26.54% of the observations with neuter sort nouns.

Regressor	Odds ratio	p
(Intercept)	0.03	< 0.001 ***
QNCASE=NOM	0.99	> 0.1 —
QNCASE=DAT	2.5	< 0.001 ***
QNDEF=1	14.48	< 0.001 ***
LOGFREQ	2.27	< 0.001 ***
QNABBR=1	0.01	< 0.001 ***

Table 2: GLM results for neuter sort nouns

There is no overdispersion (masculine φ=0.99, neuter φ=0.95), Nagelkerke's R^2 is adequate (masculine R^2=0.22, neuter R^2=0.29), and in 10-fold cross validation, the model achieves an accuracy of 0.78 for masculine and 0.76 for neuter nouns.

What do these results tell us? It should be kept in mind that the genitive construction—as mentioned above—is generally dispreferred relative to the agreement construction, and it is also historically becoming rarer. It seems clear to us that this development sharpens the division of labor between the genitive occurring in NPs with determiners such as (2) and (3b) and ones without a determiner. In Section 2 we already argued that (3b) exemplifies the odd construction in this respect. Consequently, it is becoming extinct.

The models point to a clear preference for case agreement when the quantifying noun is in a structural case (cf. QNCASE with the accusative on the low intercept and the fact that the nominative is not significantly different from it but the dative is). This is most likely due a second tendency in German morpho-syntax, namely to sharpen the distinction between the two structural cases (nominative and accusative) on the one side and the two oblique cases (dative and genitive) on the other side, while conflating differences within the two groups. The construction [N_{DAT} NP_{GEN}] with a very flat obliqueness contour between dative and genitive is thus closer to the generally preferred agreement construction than [N_{NOM} NP_{GEN}] would be, and thus it occurs relatively more often.

Furthermore, the genitive is generally associated with a more distinguished style, partly because it is already extinct in many dialects. This most likely accounts for the fact that abbreviated quantifying nouns (such as *kg*), which are more typical of a technical style, do not go well with the genitive construction (cf. QNABBR). The observed effect is thus, in our view, just an indirect marker of a register- or style-specific preference.

A high (log-)frequency of the sort-denoting noun also seems to favor the choice of the genitive construction, maybe by way of entrenchment of highly frequent collocations like "bottle of beer", "glass of wine" etc., which preserve the older genitive construction better than less frequent constructions.

The strongest effect, however, is also the most difficult to explain: why should a definite determiner on the outer head noun influence the choice of the case of the embedded noun? While the interpretation of this should be semantic in nature, we have found the semantic literature on partitives and similar phenomena to be of little help, and we leave this point open for future analyses (possibly with a more fine-grained look at single determiners) and discussion.

Further results not presented here for space reasons include a differentiated look at the count/mass distinction in the embedded NP and Generalized Linear Mixed Models (GLMMs) with the noun lemma of

100 Since the relevant distinction between the agreement and genitive construction is undetectable with genitive head nouns, observations where the head noun was in the genitive were not taken into account.

the sort-denoting noun as a random effect. These models have an improved prediction accuracy, which points towards item-specific preferences, which might also be attributed to differences in style or register.

References

Anttila, A. and Fong, V. 2000. "The Partitive Constraint in Optimality Theory". *Journal of Semantics* 17(4): 281–314.

Barker, C. 1998. "Partitives, double genitives and anti-uniqueness". *Natural Language and Linguistic Theory* 16(4): 679–717.

Baroni, M. and Bernardini, S. and Ferraresi, A. and Zanchetta, E. 2009. "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora". *Language Resources and Evaluation* 43(3): 209–226.

Eisenberg, P. (author) and Fuhrhop, N. (collaborator). 2013. *Grundriss der deutschen Grammatik: Das Wort*, Stuttgart: Metzler.

Fahrmeir, L. and Kneib, T. and Lang, S. and Marx, B. 2013. *Regression – Models, Methods, and Application*. Berlin: Springer.

Hentschel, E. 1993. "Flexionsverfall im Deutschen? Die Kasusmarkierung bei partitiven Genitiv-Attributen". *Zeitschrift für Germanistische Linguistik* 21(3): 320–333.

Vos, H. M. 1999. *A grammar of partitive constructions*. Tilburg : Tilburg University.

The notion of *Europe* in German, French and British election manifestos. A corpus linguistic approach to political discourses on Europe since 1979

Ronny Scholz
University of Warwick
r.scholz@warwick.ac.uk

In this talk I will present the results of a comparative study on the notion of *Europe* in French, German and British political discourse. This talk is taken from a larger project entitled "The discursive legitimation of the European Union" (Scholz 2010). The starting point for this research was the idea that with the political influences originating from globalisation and Europeanisation there is a need for the communication of the same transnational phenomena to be integrated into the different national political contexts. By that I mean that for instance political decisions taken on a global or European level having an appreciable impact on the national level need to be explained in national political discourses according to the particular political culture of each country containing certain rules and narrative structures. If this is true, the question is how far the representations of the same transnational phenomenon concur or differ from each other in the different national political discourses. According to this idea, the original project aimed at a detailed investigation of the linguistic construction and discursive negotiation of the notion *Europe* by analysing the linguistic use of the signifier *Europe* in different historical and cultural contexts. It followed the hypothesis that in discourses Europe only exists as floating signifier (Laclau and Mouffe 1985) which changes its meaning according to its co-text and context. That means during the temporal period of a discourse, which is understood here as a thematic debate over a certain span of time, the signifier *Europe* acquires different referential meanings depending on its co-text and context. The discursive negotiations around the signifier *Europe* are understood as having an impact on belief in the legitimacy (Weber 1922) of the European political influence in each of the three countries investigated. A comparative analysis of different political discourses in Europe sheds light into the different language practices which draw on and reconstruct different political cultures. By looking at the signifier *Europe* in national election manifestos for the European Parliament we can see how far its linguistic construction in different national discourses overlaps, and what discursive

particularities exist in each country.

This study draws on an analysis of three German, French and British corpora consisting of election manifestos of all political parties that have been elected to the European Parliaments at least once since 1979. Each corpus has a volume of around 300,000 tokens and contains around 50 texts originating from national political parties ranging from the right wing to the left. For this study I used contrastive methods such as multifactor analysis, descending hierarchy classification, keywords and collocation which have been developed in French discourse analysis since the 1970s under the label of 'lexicometrics'. I identified keywords and investigated them with regard to the differences in the political culture in all three countries. On the basis of an the analysis of how the linguistic sign *Europe* is used, inferences are made about the more general understanding of the notion *Europe* in different political cultures.

In most cases corpora for lexicometric studies are composed specifically to analyse a certain thematic discourse from both a synchronic and a diachronic perspective. Corpora often compiled are of speeches of state presidents and governments, texts produced by political parties or in historical periods of radical change in public discourse and society such as 1848 or Mai 1968 (Labbé 1998; Labbé and Monière 2003; Leblanc 2010; Mayaffre 2004; Tournier 2007, 2012; Tournier et al. 2010). Based on standardised quantifying methods lexicometrics constitutes a quantitative heuristic approach (for an introduction see Kuck and Scholz 2013; Lebart et al. 1998; Scholz and Mattissek 2014; Tournier 1975). The different quantifying methods thus alienate the researcher in the first instance from the textual material and help to find new results based on statistical measures instead of hermeneutic analysis of meaningful sequences. The "interpretative reflex" is deferred to a later point in research, when we have found linguistic elements that are salient in a certain 'partition' (Tournier 1993). Partitions normally refer to context variables of the text production that are external to the textual corpus. Heuristic methods such as multifactor analysis or descending hierarchy classification take each word token of a given text-corpus into account by contrasting different partitions. They are introduced into the corpus by the researcher according to its research interests and hypotheses. In general, lexicometric studies do not make use of reference corpora but contrast the linguistic material of different authors/speakers (like politicians or newspapers) or/and time periods (date, year of a publication, and larger time periods).

The analysis presented in this talk shows that *Europe* is used in all three countries in connection with social values as democracy, freedom and peace. Furthermore, it shows that when relating *Europe* to the nation state the political discourses seem to follow a particular way of representing the sovereignty of each individual nation and its independence in relations to the politics of the European Union. Collocation analysis shows that *Europe* has a certain core meaning in all three corpora analysed, but at the same time it acquires large variation in its meaning. This characteristic allows for Europe to be represented both as a community of values (e.g. the originator of humanism) and a political entity (e.g. the European Union) at the same time. If *Europe* simultaneously means a community of values and a political entity, then the meaning of community of values can easily shift to embrace the political entity of the European Union. Thus, the European Union can be represented as a community of values which is in contradiction with her role as a technocratic operator with particular economic interests. In the study this values-driven representation of the European Union was considered as part of an ideology which supports the policy of the European Union. This ideology bases its argument on "common social values" in order to convince the citizens of the political legitimacy of the European Union, which above all is not a project for establishing certain social values but rather for pursuing economic success in a certain part of the world.

Furthermore, by analysing the co-texts in which *Europe* occurs along with a name referring to one of the three national states, I found that using the term *Europe* facilitates a representation of national states in which their sovereignty and the competencies of their political actors seem to be untouched by EU politics. This representation was considered as another dimension of an EU ideology which legitimises its political influence because, at least until the ratification of the Lisbon Treaty (2009), national political actors were to at all disposed to participate in EU policy making or to resist the ratification of European Community directives into national law. In this regard, all three discourses seem to follow different strategies to represent the sovereignty of the national state. In this sense, in the British corpus *Europe* is very often used in negations and conditional clauses. Therefore its referential meaning stays weak throughout the research period; whereas the political parties and the traditional territory for their political action are represented with direct and affirmative wording. In this corpus, Europe appears as a very weak political subject, and in contrast the United Kingdom and its political actors appear as the only legitimate political actors for all political decisions. The strategy in the co-texts of the French corpus is quite different. Here, Europe is not only represented with all the powers of

a political subject, but it is also represented as the means of realising French political economic and technological interests in world. Therefore throughout the research period *Europe* is conceptualised much more strongly than in the British corpus. The political interests of *France* and *Europe* seem to converge and conflicts between European and French political actors seem not to exist. In both British and French texts, *Europe* seems important for the foreign policy of both countries which, however, is put into action only by politicians from the national political level. Similar to the French corpus the concept of Europe is considerably elaborated in the German corpus. However, *Europe* does not seem to have so much impact on German national politics because its conceptualisation stays limited to the idea of a community of humanist values.

References

Kuck, K. and Scholz R. 2013. „Quantitative und qualitative Methoden der Diskursanalyse als Ansatz einer rekonstruktiven Weltpolitikforschung. Zur Analyse eines internationalen Krisendiskurses in der deutschen Presse.“ In U. Franke and Roos U. (eds.) *Rekonstruktive Methoden der Weltpolitikforschung. Anwendungsbeispiele und Entwicklungstendenzen*. Baden-Baden: Nomos, 219-270.

Labbé, D. 1998. *La richesse du vocabulaire politique: de Gaulle et Mitterrand*. Paris: Champion.

Labbé, D. and Monière D. 2003. *Le discours gouvernemental. Canada, Québec, France (1945 - 2000)*. Paris: Honoré Champion.

Laclau, E. and Mouffe Ch. 1985. *Hegemony and Socialist Strategy. Towards a Radical Democratic Politics* London: Verso.

Lebart, L., Salem, A. and Berry, L. 1998. *Exploring textual data*. Dordrecht: Kluwer.

Leblanc, J.-M. 2010. „Le style Sarkozy à l’aune du rituel politique et discursif.“ *La matière et l’esprit* 17/18: 77-112.

Mayaffre, D. 2012. *Nicolas Sarkozy - Mesure et démesure du discours (2007-2012)*. Paris: Presses de Sciences Po.

Mayaffre, D. 2004. *Paroles de Président, Jacques Chirac et le discours présidentiel sous la Ve République*. Paris: Honoré Champion.

Scholz, R. 2010. *Die diskursive Legitimation der Europäischen Union. Eine lexikometrische Analyse zur Verwendung des sprachlichen Zeichens Europa/Europe in deutschen, französischen und britischen Wahlprogrammen zu den Europawahlen zwischen 1979 und 2004*. Unpublished (binational) PhD thesis, University of Magdeburg and Paris-Est. Available online at http://edoc.bibliothek.uni-halle.de/servlets/DocumentServlet?id=9670

Scholz, R. and Mattissek A. 2014. „Zwischen Exzellenz und Bildungsstreik. Lexikometrie als Methodik zur Ermittlung semantischer Makrostrukturen des Hochschulreformdiskurses.“ In M. Nonhoff et al. (eds.), *Diskursforschung. Ein interdisziplinäres Handbuch. Band 2: Methoden und Analysepraxis. Perspektiven auf Hochschulreformdiskurse*. Bielefeld: transcript, 86-112.

Tournier, M. 2007. *Les mots de mai 68*. Toulouse: Presses universitaires du Mirail.

Tournier, M. 1993. *Lexicometria - Séminaire de lexicométrie*. Lisbonne: Universidade Aberta.

Tournier, M. 1975. *Un vocabulaire ouvrier en 1848. Essai de lexicométrie. Quatre volumes multicopiés*. Saint-Cloud: École Normale Supérieure.

Tournier, M. et al. 2010. *Des noms et des gens en république (1879-1914)*. Paris: L’Harmattan.

Weber, M. 1922. *Wirtschaft und Gesellschaft* Tübingen: Mohr.

The phraseological profile of general academic verbs: a cross-disciplinary analysis of collocations

Natassia Schutz
Université catholique de Louvain
natassia.schutz@uclouvain.be

1 Introduction

The applied corpus-based research aiming to describe the linguistic features of general academic writing has witnessed a shift of interest from individual vocabulary items towards larger phraseological units. This results from the findings of various studies demonstrating, for instance, the challenge multi-word units (MUWs) represent for learners (e.g. Nation 2001), the importance of MWUs for academic proficiency (e.g. Hyland 2008), and the salience and systematic functionality of some MWUs (e.g. Biber and Barbieri 2006). In line with this research trend, this paper aims to analyze the collocational patterns of general academic verbs and identify those that cut across disciplinary boundaries so as to provide EAP learners with detailed and useful information on the phraseological profile of academic verbs.

Among the few studies examining collocations in general academic English are Durrant (2009) and Ackermann and Chen (2013). Both studies set out to build pedagogically useful lists of collocational patterns. To this aim, they automatically extracted, on the basis of inferential statistics, the collocational patterns that are typical of general academic writing. While Durrant's collocation list includes both lexical (e.g. *significantly different*) and grammatical collocates (e.g. *show that*), Ackermann and Chen's list is restricted to lexical collocates.

In this paper, we intend to analyze collocational patterns from a different perspective: rather than adopting a general approach to academic writing, we focus exclusively on academic verbs and compare their collocational patterning across academic disciplines. We believe that it is also important to take such an approach to academic phraseology as learners have been shown to struggle with the lexico-grammatical patterning of general academic verbs (e.g. Granger and Paquot 2009). By drawing a cross-disciplinary comparison, we hope to be able to distinguish the collocates that can be considered as common-core from those that are discipline-specific, and thereby, determine the weight that cross-disciplinary collocates represent in comparison with discipline-specific collocates. The results of such an analysis could thus provide useful information on cross-disciplinary collocational patterning for general EAP courses and textbooks.

2 Data, verb selection and collocation extraction

This study makes use of a 3 million-word corpus containing research articles in business, linguistics and medicine, viz. the Louvain Corpus of Research Articles (LOCRA)[101].

In order to be considered as potential general academic verbs, the verbs occurring in LOCRA had to be identified as either highly frequent[102] or key[103] in all three disciplines. The novelty of this selection procedure is that it takes into account different types of verbs that have hitherto never been considered together; academic vocabulary lists have either been based on the analysis of traditional frequencies (e.g. Coxhead 2000) or the analysis of inferential statistics (e.g. Paquot 2010). This method generated a list of 177 verbs (cf. Schutz 2013). To reduce this list to one that is more manageable for the purpose of the study, we focused on the top 15 academic verbs occurring in each discipline. The final verb list totaled 31 academic verbs (see Table 1), which represent a sizeable proportion (26%) of the verb tokens occurring in LOCRA.

appear, associate, base, consider, compare, describe, determine, develop, examine, express, find, follow, give, include, increase, indicate, influence, involve, make, observe, occur, perform, provide, receive, relate, report, see, show, suggest, take, use

Table 1: Short list of general academic verbs

The collocational analysis was carried out using the Word-Sketch option of the SketchEngine (Kilgarriff et al., 2004). This tool automatically extracts the collocates (using the logDice measure) of a specific node and categorizes them according to their grammatical relation with the node. To ensure the pedagogical relevance of the collocates, we set an additional frequency threshold of 5 occurrences with the node. The collocates of each of the 31 verbs in Table 1 were compared across business, linguistics and medicine. On this basis, we identified the collocates that are used across the three

[101] http://www.uclouvain.be/en-cecl-locra.html. LOCRA is a corpus currently under development at the *Centre for English Corpus Linguistics* which aims to represent expert academic discourse in several disciplines. It currently contains research articles from peer-reviewed top-rated journals in business, linguistics and medicine.

[102] To be considered as highly frequent, the verbs had to be among the top verbs that cover up to 80% of the total number of verb tokens in each discipline (cf. Schutz 2013).

[103] Key verbs are verbs which "occur with unusual frequency in a given text" when compared to a reference corpus (Scott 2001: 236). In this study, a corpus of fiction was used.

disciplines in LOCRA and those that are used in only one discipline.

3 Preliminary results

In terms of **types**, the results reveal that the mean percentage of collocates used across business, linguistics and medicine is 7%; the majority of the collocates identified for each verb are used in only one discipline (mean percentage: c. 80%). When examining the categories of collocates that are used across business, linguistics and medicine, it appears that most of them are subject collocates, object collocates and modifiers. This study zooms in on the results yielded for the subject and object collocates.

The **cross-disciplinary subject collocates** represent little in terms of types but cover a considerable proportion of the total number of subject tokens used in LOCRA: on average, they cover 30.4%, 22.8% and 35.2%, respectively, of the subject tokens occurring in business, linguistics and medicine. When compared to the subject collocates found in one discipline, it appears that these subject collocates cover either more than or as much as those that are discipline-specific. The cross-disciplinary subject collocates of the verb SHOW (e.g. *result* and *data*), for example, have a wider coverage (c. 30%) than those that are specific to linguistics (e.g. *speaker* and *excerpt*; 13%). The cross-disciplinary subject collocates of the verb INDICATE (e.g. *finding* and *result*), on the other hand, cover as much as the subject collocates that are specific to medicine (e.g. *abscissa* and *line*): they each cover c. 30% of the subject tokens used with INDICATE in medicine.

The **cross-disciplinary object collocates** cover a much smaller proportion (c. 15%) of the total number of object tokens than the cross-disciplinary subject collocates. When compared to the collocates used in one discipline only, a different trend from the one described above was found. In this case, it is the collocates used in one discipline only which either cover more than or as much as those found across the three disciplines. The object collocates used with GIVE in medicine only (e.g. *consent* and *dose*), for example, cover more (27%) than the cross-disciplinary object collocates identified in LOCRA (e.g. *rise* and *result*; 7%). The object collocates used with TAKE in business only (e.g. *action* and *measure*), on the other hand, cover as much as the cross-disciplinary collocates identified in LOCRA (e.g. *advantage* and *place*): they each cover c. 25% of the object tokens used with TAKE in business.

A more qualitative look at the cross-disciplinary subject and object collocates reveals that they all seem to be good candidates for general EAP courses as they clearly relate to the core business of research, irrespective of the discipline. The object collocates, for instance, refer to academic activities, such as the analysis and reporting of a type of relationship (see examples 1 and 2) or the description of a methodology (see examples 3 and 4).

(1) *(MED) To **compare** the **effect** of genotype on the risk of new-onset asthma, we also performed a stratified analysis.*

(2) *(BUS) The results also **show** no **difference** in the stock preferences of American-, European- and Asian-based funds.*

(3) *(LING) Bickford 1991 also **performed analyses** on signs between LSM and ASL that are articulated similarly.*

(4) *(MED) We **used** current clinical **criteria** to diagnose kidney disease instead.*

The collocates found in only one discipline, on the other hand, did not all seem to be discipline-specific. Next to the ones that are undoubtedly discipline-specific are a few collocates that could be found in other disciplines (e.g. *section/article* + DESCRIBE; see examples 5 and 6). The reason for this is that our corpus contains only three disciplines. While it appears to be sufficient to identify interesting cross-disciplinary collocates, our results show that a more diversified corpus is necessary to better discriminate the collocates that are discipline-specific from the others.

(5) *(BUS) The final **section describes** the findings, and presents a discussion of the major issues arising from the study.*

(6) *(LING) This **article describes** a novel use of underutilized recordings of moribund folklore.*

4 Conclusion

The cross-disciplinary comparison carried out in this study proved successful in identifying a large number of cross-disciplinary collocates of academic verbs that could be usefully integrated into pedagogical materials. Our results show that it is subject collocates, rather than object collocates, that tend be cross-disciplinary. Discipline-specific collocates tend to appear more as object collocates. Conducted on a larger scale, cross-disciplinary comparisons along the lines presented here highlight typical phraseological patterns of academic verbs which EAP teachers could use to raise their students' awareness as to how general academic vocabulary behaves across disciplines and in their own field of study.

References

Ackermann, K. and Chen, Y.-H. 2013. Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, 12 (4): 235–247.

Biber, D., and Barbieri, F. 2006. Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26 (3): 263-286.

Coxhead, A. 2000. A New Academic Word List. *TESOL Quarterly*, 34 (2): 213-238.

Durrant, P. 2009. Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, 28 (3): 157–169.

Granger, S. and Paquot, M. 2009. Lexical verbs in academic discourse: a corpus-driven study of learner use. In M. Charles, D. Pecorari and S. Hunston (eds.) *Academic writing: at the interface of corpus and discourse*. London: Continuum: 193-214.

Hyland, K. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27* (1): 4-21.

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. 2004. The Sketch Engine. In G. Williams and S. Vessier, (eds.) *Proceedings of the Eleventh EURALEX International Congress*: 105-116.

Nation, P. 2001. Learning vocabulary in another language. Cambridge: Cambridge University Press.

Paquot, M. 2010. *Academic vocabulary in learner writing*. London: Continuum.

Schutz, N. 2013. How Specific is English for Academic Purposes? A look at verbs in business, linguistics and medical research articles. In G. Andersen and K. Bech (eds.). *English Corpus Linguistics: Variation in Time, Space and Genre.* Amsterdam: Rodopi Publishers: 237-257.

Scott, M. 2001. Comparing corpora and identifying keywords, collocations, frequency distributions through the WordSmith Tools suite of computer programs. In M. Ghadessy, A. Henry and L. Roseberry (eds.). *Small Corpus Studies and ELT*. Amsterdam: John Benjamins: 47-67.

Life-forms, Language and Links: Corpus evidence of the associations made in discourse about animals

Alison Sealey
Lancaster University
a.sealey@lancaster.ac.uk

1 Introduction and rationale

This paper draws on findings emerging from a three-year research project, funded by the Leverhulme Trust, into the characteristics of discourse about animals[104]. The overall aim of this project is to investigate patterns in the way animals are discursively represented, not only because this may be intrinsically interesting, but also because of the light it can potentially shed on the relationship between discourse and reality. That is: a wide range of discourse analytic studies (including some from CDA perspectives, some using corpus methods and some both) have explored how various social groups are represented in language (e.g. Baker 2006; Baker et al. 2008; Baker et al. 2013; Caldas-Coulthard and Moon 2010; Gabrielatos and Baker 2008; Litosseliti and Sunderland 2002; Partington 2004; Partington et al. 2004). Such research must engage with the possibility of reflexivity, in that the people described may themselves respond to – and sometimes contribute to – these discursive representations. This project, by contrast, focuses on language about living, sentient beings that, since they lack human linguistic resources, do not participate directly in the production of discourse. Thus, patterns in the language used to represent animals and what they do are a product of both the objective characteristics of the creatures and the way discourse is used to convey human perceptions of and stances towards them.

Animals feature in human experience and discourse as: objects of observation, study or entertainment (in the wild, in laboratories, in zoos); companions; tools (for transport and/or work); commodities (for meat, other edible products, fur and clothes), competitors (with each other and with humans, in sport, as quarry in hunting, racing, fighting) and out of place (pests / vermin) (see DeMello 2012; Herzog 2010; Ingold 1988). These are not mutually exclusive categories: creatures hunted for sport, such as game birds or fish, may then be eaten; creatures regarded as pests or vermin may be executed clinically (e.g. by fumigation) or hunted down in sporting rituals (e.g. foxes).

[104] *People, products, pests and pets: the discursive representation of animals* (RPG 2013 063)

Likewise, there are often no neat divisions between kinds of animal and orientations towards them: a dog may be treated as a pet or made to compete in fights, used for guarding the home or acting as a blind person's 'sight'. 'Animals', then, and people's experience of and beliefs about them, are quite heterogeneous, and the composition of our corpus reflects this heterogeneity.

2 Data

The data for our project comprises a corpus of texts (both writing and transcribed speech), produced between 1995 and 2015, which have animals as a central theme. In this corpus we have included items from newspapers, commentaries from television broadcasts about 'wildlife', literature produced by organisations campaigning on animal-related issues, food product labels, etc. In addition, we have interviewed 15 people who are centrally involved in the production of such texts, including broadcasters, scientists, environmentalists, animal welfare campaigners, farmers' representatives, etc. The main focus of these interviews, which have been transcribed and now comprise a further sub-set of our corpus, is the discourse producers' views on what constitutes the optimal language to achieve their purposes. A third dataset within our corpus is transcriptions of the focus groups we have conducted, some with various interest groups, others with members of the general public, to ascertain their responses to a series of selected texts from other parts of the corpus.

3 Research questions

Our research examines how language choices relate to particular scientific, philosophical, ethical, popular and practical stances towards animals, seeking answers to such questions as:

- How are animals in the corpus represented by the language used, and how does this vary with genre and purposes?
- What kinds of description are associated with different kinds of animal?

4 Findings

In the process of answering these broad questions, we have been faced with the need to categorise the different kinds of animals for which naming terms feature in our corpus. We have consulted people with expertise in this area, including biologists, ethologists and philosophers. The taxonomies they use relate to the – sometimes contrasting – purposes behind the classification (see Dupré 2001; 2002; 2012), and the paper focuses on some patterns in the discursive categorisations of animals that are emerging from the analysis of the discourse contained in our corpus. In this sense, the corpus-assisted discourse analysis contributes to our understanding of the range of ways in which people conceptualise these denizens of the natural world, and provides an indication of the links between perceptual, attitudinal, practical and linguistic classifications.

References

Baker, P. 2006. *Using Corpora in Discourse Analysis*, London: Continuum.

Baker, P., Gabrielatos, C. and McEnery, T. 2013. *Discourse Analysis and Media Attitudes: the representation of Islam in the British press*, Cambridge: Cambridge University Press.

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyzanowski, M., McEnery, T. and Wodak, R. 2008. 'A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press', *Discourse and Society* 19(3): 273 - 306.

Caldas-Coulthard, C. and Moon, R. 2010. 'Curvy, hunky, kinky': Using corpora as tools for critical analysis**.** *Discourse & Society* 21: 99-133

DeMello, M. 2012. *Animals and Society: an introduction to human-animal studies*. New York, Columbia University Press.

Dupré, J. 2001. 'In defence of classification', *Studies in History and Philosophy of Biological and Biomedical Sciences* 32: 203–219.

Dupré, J. 2002. *Humans and Other Animals*, Oxford: Clarendon Press.

Dupré, J. 2012. *Processes of Life*, Oxford: Oxford University Press.

Gabrielatos, C. and Baker, P. 2008. 'Fleeing, sneaking, flooding: a corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996-2005', *Journal of English Linguistics* 36(5): 5 - 38.

Herzog, H. 2010. *Some We Love, Some We hate, Some We Eat: why it's so hard to think straight about animals*: Harper Perennial.

Ingold, T. (ed.) 1988. *What is an Animal*: Unwin Hyman.

Litosseliti, L. and Sunderland, J. 2002. *Gender Identity and Discourse Analysis*. Amsterdam: John Benjamins

Partington, A. 2004. 'Corpora and discourse, a most congruous beast', in A. Partington, J. Morley and L. Haarman (eds.) *Corpora and Discourse*, pp. 11 - 20. Berlin: Peter Lang.

Partington, A., Morley, J. and Haarman, L. (eds.) 2004. *Corpora and Discourse*, Berlin: Peter Lang.

Teaching Near-Synonyms More Effectively -- A case study of 'happy' words in Mandarin Chinese

Juan Shao
Xi'an Jiaotong University
University of Liverpool
juan.shao@ liverpool.ac.uk

The use of corpora in language teaching has been gaining increasing prominence in the last two decades. A great number of corpus-related (corpus-based, corpus-driven and corpus-assisted) research studies have contributed to the advancement in language pedagogy, in particular teaching English as a second/foreign language (TESOL). The topics range from compiling corpus-driven dictionaries for learners, designing supplementary teaching materials as well as textbooks and using corpora in the classroom to analysing learner language, conducting comparative study between first and target language and teaching ESP/EAP. Most of the research has concentrated on English and some on European languages such as French and Spanish; few studies, however, have been done on Chinese Mandarin even though the last decade has witnessed a boom in learning Chinese as a second/foreign language across the world.

A number of research studies have been conducted on linguistic behaviours of lexis, phraseology, pattern grammar, n-grams (Sinclair 2004; Stubbs 2007; Granger & Meunier 2008), and the findings have been applied in English language pedagogy. Chinese linguistics remains a field less explored. Xiao and McEnery (2010) conducted a brilliant contrastive study between English and Mandarin Chinese focusing on tense and aspect, which not only enriches linguistic description but also provides a potential reference point for Chinese teaching and learning. However, much work needs to be done in the area of teaching Chinese as a second/foreign language including corpus-based linguistic description and its pedagogic applications.

This study focuses on one important and difficult aspect in teaching which has scarcely been explored, namely distinction in near-synonyms. Despite its importance and intricacy, synonymy has not garnered the scholarly attention it deserves until quite recently (Divjak 2006, Edmonds & Hirst 2002, Taylor 2002). 'Because of their subtle nuances and variations in meaning and usage, synonyms offer an array of possible word choices to allow us to convey meanings more precisely and effectively for the right audience and context' (Liu and Espino 2012); thus how to choose the most appropriate one from a list of synonyms for the right audience and context constantly frustrates language learners and also poses a problem for teachers. The aim of the study is to explore ways to effectively explain how near-synonyms are distinguished based on corpus exploration. The findings will help the students to make better decisions in choosing the most appropriate word or phrase for the right audience and context.

Lexical priming (Hoey 2005) seems to provide an excellent theoretical and practical framework for distinguishing members of a pair/group of synonyms. Lexical priming has universal application and its applicability to Chinese has been demonstrated in previous studies (Hoey and Shao forthcoming; Shao in preparation). As Hoey (2005) points out 'synonyms differ in respect of the way they are primed for collocation, semantic associations, colligations and pragmatic associations'. The first aim of the study is to explore linguistic behaviour of Mandarin Chinese 'happy' words: 高兴 (*gāo xìng*, happy/happily)[105], 快乐 (*kuài lè*, happy/glad) and 开心 (*kāi xīn*, happy/glad). A detailed analysis is conducted to investigate the similarities and differences of the words in terms of collocation, semantic association, colligation and pragmatic association.

In addition, it has been widely acknowledged that words and phrases behave differently in various genres and text types. Language learners may make mistakes or confuse readers by choosing the most frequently used common word without considering the genre or the purpose of their writings. The second aim of the study is to find out the frequency of use and related collocations of near synonyms in different sub-corpora.

Finally, language transfer from L1 may also influence how learners use the target language. A common goal in language learning is to write and speak in a way as much like native speakers as possible. Hoey (2005) has pointed out that by studying corpora we can get clues into how people are primed and how they tend to use particular words and phrases. The final aim of the study is to look at how English and Chinese 'happy' words are similar and different in terms of collocation, semantic association, colligation, and pragmatic association. The comparison between the two languages may show why learners make certain mistakes in the target language and thus indicate ways of improving the learning of Chinese near-synonyms.

Therefore my research questions are: (1) How are Chinese words meaning 'happy' primed in terms of

105 The Chinese is given first in character form, then in Pinyin, followed by a free translation.

collocation, semantic association, colligation and pragmatic association? (2) How are these 'happy' word distributed in different genres and text types? (3) Is there a potential link between the way English speakers are primed with respect to words meaning 'happy' in English and the way they use similar words in Mandarin Chinese?

To tackle these questions, the Lancaster Corpus of Mandarin Chinese (LCMC) was analysed with CQPweb (Hardie 2012) and FLOB was analysed with the Sketch Engine (Kilgarriff 2008) due to the accessibility of the corpus. Since LCMC is designed as a Chinese match for the FLOB and FROWN corpora of modern British and American English respectively, the comparability of the corpora may offer more reliable findings.

The findings concerning the first question are as follows. There is a long list of collocates of 高兴 (*gāo xìng*) including 地 (*de*, adverb suffix), 很 (*hěn,* very), 心里 (*xīn lǐ*, in the heart), 非常 (*fēi cháng*, very), 十分 (*shí fēn*, very), 我 (*wǒ*, I), 太 (*tài*, too), 听 (*tīng*, listen), 不 (*bù*, not), 了 (*le*, functional word), 说 (*shuō*, speak/talk), 他 (*tā*, he), 她 (*tā*, she), 就 (*jiù, functional word*). As for 快乐 (*kuài lè*), collocates include 祝 (*zhù*, bless), 生日 (*shēng* rì, birthday), 感觉 (*gǎn jué*, feel), 你 (*nǐ*, you), 不 (*bù*, not), 人 (*rén*, person), 的 (*de*, adjective suffix), 是 (*shì*, BE). 开心 (*kāi* xīn) yields fewest collocates, namely 地 (de, adverb suffix) and 的 (de, adjective suffix).

By looking at the collocates of 高兴 (gāo xìng), it is not difficult to identify semantic sets. Firstly, 高兴 (*gāo xìng*) co-occurs with intensifiers such as 很 (*hěn*, very), 非常 (*fēi cháng*, very), 十分 (*shí fēn*, very), and 太 (*tài*, too). Secondly, verbs denoting sensory experiences appear in another semantic set, including 听 (*tīng*, listen) and 说 (*shuō*, speak/talk). Last, personal pronouns such as 我 (*wǒ*, I), 他 (*tā*, he) and 她 (*tā*, she) form a third semantic group.

With 快乐 (*kuài lè*), a restricted domain can be identified from the collocates, that is birthday celebration. 祝 (*zhù*, bless) and 生日 (*sheng rì*, birthday) are included. Of interest is the situation of collocate 你 (*nǐ*, you). One may argue that it cannot be categorised into the current semantic group, however, the examination of the concordances show 3 out 4 hits are related to the topic as it is used in the structure 祝你生日快乐 (*zhù nǐ shēngrì kuài lè,* wish you a happy birthday).

The colligational analysis of the three 'happy' words focuses on the co-occurrence of the Chinese suffixes 地 (*de, adverb suffix*) and 的 (*de*, adjective suffix). The first word 地 (*de, adverb suffix*) in the collocation list of 高兴 (*gāo xìng*) shows that 高兴 (*gāo xìng*) usually co-occurs with the suffix 地 (*de, adverb suffix*) to form an adverb to modify another verb. In addition, Hoey (2005) emphasises that colligation includes the avoidance of certain grammatical patterns and functions. The negative collocation of the suffix 的 (*de, adjective suffix*) is an example of this. On the other hand, 快乐 (*kuài lè*) is usually used as an adjective and consequently occurs with suffix 的 (*de, adjective suffix*) to modify nouns. Both 地 (*de, adverb suffix*) and 的 (*de, adjective suffix*) occur in the collocation list of 开心 (*kāi xīn*), which indicates that 开心 (*kāi xīn*) functions with equal facility as adjective and adverb.

One significant pragmatic association that needs to be mentioned here is that of 快乐 (*kuài lè*) with Birthday Celebration; as Hoey (2005) states, semantic association is linked to pragmatic association.

To address the second question, the three 'happy' words 高兴 (*gāo xìng*), 快乐 (*kuài lè*) and 开心 (*kāi xīn*) are distributed differently in terms of frequency and dispersion. 高兴 (*gāo xìng*) has the highest frequency with 123 matches (122.83 hits per million) and dispersion (82 texts out of 500). 快乐 (*kuài lè*) ranks the second with 26 hits (25.96 per million) across 20 (out of 500) texts and 开心 (*kāi xīn*) only occurs 9 times (8.99 per million) in 7 texts.

The results of analysis of the English data were compared with those for Chinese and the difference between the groups offer a potential explanation of the difficulty in using near-synonyms in the target language. One limitation of the study is that without analysis of the learner's performance (for example in speaking and writing) it is impossible to find out whether there exists a direct link between the learner's primings in the first and target languages. Research on interlanguage may provide more reliable evidence of priming transfer.

Despite of these limitations, this study provides some indications of the way that corpus-based study on Chinese near-synonyms and may provide insights into better ways of teaching Chinese as a second/foreign language. Further research needs to be conducted into other near-synonymous words and phrases as well as into the interlanguage of English-speaking Chinese language learners.

References

Divjak, D. 2006. "Ways of intending: Delineating and structuring near synonyms." In S. Th Gries & A. Stefanowitsch (Eds.), *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*. Berlin/New York: Mouton de Gruyter, 19–56.

Edmonds, P. & Hirst, G. 2002. "Near synonyms and lexical choice." In *Computational Linguistics*, 28 (2), 105-144.

Granger, S. & Meunier, F. (Eds.) 2008. *Phraseology: An Interdisciplinary Perspective*. Amsterdam: John Benjamins.

Hardie, A. 2012. “CQPweb -combining power, flexibility and usability in a corpus analysis tool”. In *International Journal of Corpus Linguistics* 17 (3), 380–409. John Benjamins Publishing Company.

Hardie, A. 2012. CQPweb. https://cqpweb.lancs.ac.uk

Hoey, M. 2005. *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Hoey, M. and Shao, J. (forthcoming). “Lexical Priming: The Odd Case of a Psycholinguistic Theory that Generates Corpus-linguistic Hypotheses for both English and Chinese.” In B. Zou, M. Hoey, S. Smith (Eds), *Corpus Linguistics in Chinese Contexts*. Palgrave Macmillan.

Kilgarriff, A. 2008. The Sketch Engine. https://the.sketchengine.co.uk

Liu, D & Espino, M. 2012. “Actually, Genuinely, Really, and Truly: A corpus-based Behavioral Profile study of near-synonymous adverbs.” In *International Journal of Corpus Linguistics* 17:2 198–228. John Benjamins Publishing Company.

Sinclair, J. 2004. *Trust the Text. Language, Corpus and Discourse*. London: Routledge

Stubbs, M. 2007. “Quantitative data on multiword sequences in English: The case of the word world”. In M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert (Eds.), *Text, Discourse and Corpora*. London: Continuum, 163–190.

Taylor, J. R. 2002. “Near Synonyms as Co-extensive Categories: ‘High’ and ‘tall’ revisited”. In *Language Sciences*, 25 (3), 263–284.

Xiao, R. and McEnery, T. 2010. *Corpus-based contrastive studies of English and Chinese*. New York/London: Routledge.

Approaching genre classification via syndromes

Serge Sharoff
University of Leeds
s.sharoff@leeds.ac.uk

1 Introduction

Large representative corpora need basic information about their composition in terms of genres. However, getting a suitable set of genre labels is surprisingly difficult. The major corpora disagree with respect to their genre inventories: 15 categories of the Brown-family corpora (Kučera and Francis, 1967), 70 categories in the BNC (Lee, 2001), to more than 4,000 genre labels in studies on text typology (Adamzik, 1995). Nevertheless, the task of classifying traditional corpora was reasonably straightforward because their documents came from a relatively small number of well controlled sources. The task of providing a basic genre classification of Web corpora is considerably more difficult, especially when we consider a degree of genre hybridism (Santini et al., 2010): the authors of such documents are less controlled by the institutional gatekeepers and have freedom in choosing from a wide set of genre conventions. Even in a well-controlled corpus, many texts do not get a label, for example, `W.misc` is the most frequent genre category in the BNC.

A related problem in large-scale text classification concerns inter-annotator agreement. The annotators often disagree in assigning a label, especially in the case of randomly selected texts. For example, the difference between informative reporting (Category A in the Brown Corpus) and opinionated discussion (Category B) in many texts depends on the perception of the annotator, especially when we apply the Brown Corpus categories to classifying the Web, because Web texts often blend them in various proportions. In the end, chance corrected agreement measures for such categories report values below the accepted reliability thresholds, e.g. Krippendorff’s α=0.51 for ‘reporting’ (Sharoff et al., 2010). This level of disagreement means that inferences drawn from texts, which have been classified with labels like ‘reporting’ vs ‘discussion’, are not statistically significant.

2 Annotation scheme

This paper reports an experiment in annotation of Web-based corpora using a small number of Functional Text Dimensions (FTDs), so that a text is described by the degree of its similarity to more prototypical texts using such questions as:

A8: hardnews To what extent does the text appear to be an informative report of recent events? (For example, a news item).

A1: argum To what extent does the text contain explicit argumentation to persuade the reader? (For example, an editorial)

A17: eval To what extent does the text evaluate a specific entity by endorsing or criticising it? (For example, by providing a product review)

The three dimensions above respectively correspond to the Categories A, B and C in the Brown Corpus. A text can present variation in each dimension, for example, by containing more reporting and less argumentation or vice versa. This can lead to possible changes in annotations of established corpora. For example, in addition to having purely reporting texts in Category A, some texts in the Brown Corpus also contain a fair amount of argumentation:

> *The most positive element to emerge from the Oslo meeting of North Atlantic Treaty Organization Foreign Ministers has been the freer, franker, and wider discussions, animated by much better mutual understanding than in past meetings. This has been a working session of an organization that, by its very nature, can only proceed along its route step by step and without dramatic changes…*("NATO Welds Unity" The Christian Science Monitor, A04)

In the proposed annotation scheme this text gets `A1: 1, A8:2` indicating that it is partially argumentative.

3 Inter-annotator agreement

Inter-annotator agreement can be assessed along each of these dimensions using the Krippendorff's α value (Krippendorff, 2004), which measures the ratio between disagreement within annotations for the same text vs overall disagreement. The value of $\alpha=1.0$ indicates total agreement, $\alpha \geq 0.80$ suggests reliable data, while the value of $\alpha \geq 0.67$ is needed to support acceptable reliability judgements (Krippendorff, 2004).

Double annotation of 250 texts from a multilingual corpus (Forsyth and Sharoff, 2014) shows that the level of agreement is within acceptable thresholds (Table 1).

The worst offenders are the categories of texts aimed at entertainment (A5), informal communication (A6) and texts for specialists (A15). Agreement becomes 1.00 when there are no examples of respective categories in the collection being annotated.

4 From symptoms to syndromes

Even if FTDs themselves are reliable, the practice of corpus annotation assumes labels assigned to texts, primarily because we want to assess the composition of a corpus in terms of genre categories. In the FTD framework this can be modelled via common combinations FTDs, which can be treated as established genres. M.A.K. Halliday introduced the following metaphor in the context of Systemic-Functional Linguistics: "a register is a syndrome of lexicogrammatical probabilities" (Halliday, 1992, p. 68). The term 'register' in Halliday's terminology is related to the properties of language use. If 'genre' is related to the way a text functions in the community of language users, it can be also treated as a "syndrome," a recurrent combination of several FTDs. For this task in addition to the 250 multilingual texts with double annotation, 250 randomly selected texts from ukWac (Baroni et al., 2009) have been annotated by a single annotator.

The "syndrome" genres in this corpus have been determined by clustering using the pam (partitioning around medoids) method from the cluster package in R. A commonly used silhouette method (Kaufman and Rousseeuw, 2009) suggests that the optimal number of clusters in this set is about 11-14. The 12 cluster solution generates the following genres (with the number of their instances given in brackets):

A1 argumentative texts (64);
A12 advertising texts (55);
A11 personal reporting in diary-like blog entries (51);
A16 purely informative texts (51);
A20 apellative texts, e.g., inviting the reader to take part in an activity, (43);

A1+A13 argumentative texts with explicit propaganda, often coming from blog posts (38);
A8 hard news (37);
A4 fiction (34);
A14 research papers (33);
A7 FAQ-like instructions (32);
A7+A14 instructive academic texts (30);
A9 legal texts (24)

	A1	A3	A4	A5	A6	A7	A8	A9	A11	A12	A13	A14	A15
Part1	0.91	0.79	0.97	0.69	0.69	0.98	0.86	0.89	0.80	0.88	0.93	0.90	0.59
Part2	0.80	0.95	1.00	0.97	0.91	0.99	0.78	1.00	0.94	1.00	0.89	0.91	0.90

Table 1: Krippendorff's α agreement values

While the specific frequencies of categories in this list are based on data from a relatively small corpus, half of each has been collected by targeting multilingual resources (e.g., TED, UN corpus, WikiNews), the general framework has identified more typical web-genres which can be detected reliably. An interesting observation from the clustering experiment is the need to distinguish FAQ-like instructions from instructive academic texts, even when a small genre inventory is used.

With this framework in mind, the next task is to design an automatic classification model (Sharoff et al, 2010) to assign each text from a large Web corpus, such as ukWac, to a category from this list, thus determining its genre composition with a reasonable degree of accuracy. The results will be reported at the time of the conference. This will make it possible to link the text-external perspective suggested by the Functional Text Dimensions and the text-internal perspective suggested by Biber's Multi-dimensional Analysis (Biber, 1988)

References

Adamzik, K. (1995). *Textsorten – Texttypologie. Eine kommentierte Bibliographie*. Nodus, Münster.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Biber, D. (1988). *Variations Across Speech and Writing*. Cambridge University Press.

Forsyth, R. and Sharoff, S. (2014). Document dissimilarity within and across languages: a benchmarking study. *Literary and Linguistic Computing*, 29:6–22.

Halliday, M. (1992). Language as system and language as instance: The corpus as a theoretical construct. In Svartvik, J., editor, *Directions in corpus linguistics: proceedings of Nobel Symposium 82 Stockholm*, volume 65, pages 61–77. Walter de Gruyter.

Kaufman, L., and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, John Wiley & Sons.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3).

Kučera, H. and Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press, Providence.

Lee, D. (2001). Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, 5(3):37–72.

Santini, M., Mehler, A., and Sharoff, S. (2010). Riding the rough waves of genre on the web. In Mehler, A., Sharoff, S., and Santini, M., editors, *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.

Sharoff, S., Wu, Z., and Markert, K. (2010). The Web library of Babel: evaluating genre collections. In *Proc. of the Seventh Language Resources and Evaluation Conference, LRE*

Tracing changes in political discourse: the case of *seongjang* (growth) and *bokji* (welfare) in South Korean newspapers

Seoin Shin
Hallym University
seoin.shin@gmail.com

1 Introduction

This paper examines how the discourse on economic growth and welfare in South Korean newspapers has changed. Even though South Korea has achieved remarkable economic growth in the past 50~60 years, Korean people are still eager for further growth, and seem to believe the myth that greater economic growth will automatically lead to a better life for all. However, the reality has been just the opposite: the more the economy has grown, the more the gap between rich and poor has increased because growth has resulted in the welfare issue receiving little attention. Quite recently, however, a discourse on welfare has started to emerge in South Korea. The current study chronologically traces how the balance between economic growth and welfare has developed in the political discourse in South Korean newspapers. As Stubbs (2001: 215) argued, 'Repeated patterns show that evaluative meanings are not merely personal and idiosyncratic, but widely shared in a discourse community.' And Baker (2006: 14-15) has stated the advantage of a corpus-based approach to the investigation of changing discourse. Therefore, this paper contributes to modern diachronic corpus-assisted discourse studies (MD-CADS) as Partington (2011) and Partington et al. (2013) have stated.

2 Data and Methods

In order to observe the change of discourse, the newspaper articles covering the periods of five presidential elections are collected. The reason for selecting election periods is that people's political hopes and desires are most prominently voiced during these times when the presidential candidates present their vision for the country by reflecting the desires of people in their campaign, and this in turn provokes a wealth of discussion in the media.

A modern diachronic corpus was built by collecting newspaper articles for one month just prior to each election.

- 14th election: 18/11/1992~17/12/1992
- 15th election: 18/11/1997~17/12/1997
- 16th election: 19/11/2002~18/12/2002
- 17th election: 19/11/2007~18/12/2007
- 18th election: 19/11/2012~18/12/2012

The corpora are composed of two parts: one is built with the newspaper articles that contain the words *seongjang* (growth) and *baljeon* (development); the other one is built with the search terms *bokji* (welfare) and *bunbae* (distribution).

The corpus for *seongjang* (growth) and the corpus for *bokji* (welfare) are analysed separately and then compared and crosschecked. The total for the words in the corpora is as below:

Corpus	Number of articles	Number of words
seongjang (growth)	7,879	3,242,207
baljeon (development)	8,515	3,535,063
bokji (welfare)	4,996	1,861,783
bunbae (distribution)	608	365,321

In order to detect meaningful changes in the discourse, a collocation analysis, keywords analysis and key clusters analysis are carried out. The concordance lines are also inspected manually. The concordancing toolkit *AntConc* (Anthony 2014) is used for data work.

3 The characteristics of discourse on *seongjang* (growth)

The main characteristic of the discourse on *seongjang* (growth) is that the quantity of growth is always of interest. The modifiers of *seongjang* (growth) show this characteristic.

- amount of growth: numbers like *4 peosenteu seongjang* (4 percent growth), *5.7% seongjang* (5.7% growth), and so on; *peulleoseu seongjang* (plus growth), *jero seongjang* (zero growth), *maineoseu seongjang* (minus growth); *muhan seongjang* (growth with no limit)
- rate of growth: *goseongjang* (high growth), *godo seojang* (high level of economic growth); *jeoseongjang* (low growth), *choejeo seongjang* (lowest growth), *cheoso seongjang* (minimum growth)
- rapidness of growth: *gosok seongjang* (fast growth), *kwaesok seongjang* (growth in high speed), *chogosok seongjang* (superfast growth), *geup seongjang* (rapid growth)

Even though the manner of growth is described, this is mainly related to the quantitative characteristics of growth.

- manner of growth: *pokbaljeok seongjang* (explosive growth), *biyakjeok seongjang* (growth by leaps and bounds)

4 The change of discourse on *seongjang* (growth)

Several distinctive changes were detected in the

course of tracing the five presidential elections. Firstly, in 1997 attention was given to the quality of the growth. People started to talk about not only the quantity but also the quality of growth. The newly appearing modifiers show this change in the discourse.

- orientation of growth: *gyunhyeong seongjang* (balanced growth, 1997), *jisokganeunghan seongjang* (sustainable growth, 2002)

Secondly, there was introspection on the rapid growth as seen in the following phrases.

- reflection on the substantiality: *geopum seongjang* (bubble growth), *oihyeong seongjang* (growth in the outward appearance)
- reflection on the process: *abchuk seongjang* (condensed growth), *seongjang ilbyeondo* (complete devotion to growth)

The change of discourse in this aspect appeared after South Korea had experienced a severe financial crisis in 2007.

Thirdly, the growth started to be evaluated. The expression *joheun seongjang* (good growth) appeared in the 2007 election. Below are the evaluative modifiers occurring with the word *seongjang* (growth):

- positive evaluation: *joheun seongjang* (good growth), *ddaddeutan seongjang* (warm growth), *geonjeonhan seongjang* (sound growth), *gyegeup eomneun seongjang* (classless growth), *chabyul eomneun seongjang* (non-discriminated growth)
- negative evaluation: *mujabihan seongjang* (merciless growth), *bujueuihan seongjang* (careless growth), *bulgyunhyeong seongjang* (unbalanced growth)

These evaluative expressions indicate that the focus of the discussion is shifting to the orientation of growth.

Fourthly, people started to discuss how the growth was generated. The phrase *seongjang dongryeok* (growth engine) began to be used more often. From 2007, the expression *sinseongjang dongryeok* (new growth engine) was used fairly often.

5 The goal in company with *seongjang* (growth): *bokji* (welfare)

By looking at the use of *seongjang* (growth) as a binomial, the change in the desires of the people in South Korea can be detected. The words linked to *seongjang* (growth) with *and* is as follows:

- binomials: *gyeongje seongjanggwa mulga anjeong* (economic growth and price stabilization), *seongjanggwa baljeon* (growth and development); *gyeongje seongjanggwa goyong* (economic growth and employment), *gyeongje seongjanggwa bokji* (economic growth and welfare), *seongjanggwa bunbae* (growth and distribution); *seongjanggwas gaehyeok* (growth and reformation), *gyeongje seongjanggwa gyeongje minjuwha* (economic growth and economic democratization)

In 1992, the most frequent pair was *gyeongje seongjanggwa mulga anjeong* (economic growth and price stabilization); in 2002, *seongjanggwa bunbae* (growth and distribution) became frequent; in 2012, *seongjanggwa bunbae* (growth and distribution) became slightly less frequent but *gyeongje seongjanggwa bokji* (economic growth and welfare) became more frequent. This change in the binomials shows the formation of discourse on welfare. When people talked about *seongjanggwa bunbae* (growth and distribution), their discussion was rather abstract and theoretical. However, when they expressed their desire with the word *bokji* (welfare), the political discussion became more concrete and substantial.

The crucial turning point in the discourse can also be detected in another way: the expression of *bunbaereul tonghan seongjang* (the growth through the distribution of wealth) is introduced in 2002. The 2002 election saw the first presidential candidate from the radical left-wing labor party and not only growth but also welfare came into the political discourse in earnest at this time.

6 Conclusion

Tracing the political discourse during the periods of presidential elections has shown the change in the balance between economic growth and welfare in South Korea. The corpus-based approach to political discourse on growth provides authentic evidence of this change. It has discovered that there was not only a shift in the perspective on growth itself but also a change of focus from growth to welfare.

A further study will analyse *bokji* (welfare), and will show the change in political discourse on welfare: from *seonbyeol bokji* (selective welfare) to *bopyeon bokji* (universal welfare).

References

Anthony, L. 2014. *AntConc* (Version 3.4.3w) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software/antconc/.

Baker, P. 2006. *Using Corpora in Discourse Analysis*.

London: Continuum.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., McEnery, T. and Wodak, R. 2008. "A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press". *Discourse and Society* 19(3), 273-306.

Baker, P. and McEnery, T. 2005. "A corpus-based approach to discourse of refugees and asylum seekers in UN and newspaper texts". *Journal of Language and Politics* 4, 197-226.

Bond, M. and Scott, M. 2010. (eds.) Keyness in Texts, John Benjamins Publishing Company.

Fairclough, N. 1989. *Language and Power*. London: Longman.

Fairclough, N. 2003. *Analysing Discourse: Textual Analysis for Social Research*. London: Routledge.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Mahlberg, M. 2007. "Clusters, key clusters and local textual functions in Dickens, " *Corpora*, Vol.2, No.1, 1-31.

Partington, A. 2011. Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS) on UK newspapers: an overview of the project, *Corpora* 5:2, 83-108.

Partington, A., Duguid, A. and Taylor, C. 2013. *Patterns and Meanings in Discourse: Theory and practice in corpus-assisted discourse studies (CADS)*, John Benjamins Publishing Company.

Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.

Stubbs, M. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell

Analyzing the conjunctive relations in Turkish and English pedagogical texts: A Hallidayan approach

Meliha Simsek
Mersin University
malliday@gmail.com

1 Introduction

Among Halliday's (Halliday and Matthiessen 2004) four ways of creating cohesion in the text, conjunction is known as the most demanding and developmentally the most difficult for the child reader, and necessitates more learner support than reference, ellipsis, and lexical cohesion (Gamble and Yates 2002). This is because conjunction forms semantic relationships between different parts of discourse, not constrained by clause boundaries, and depends on the logical progression of events more than the succession of linguistic items (Halliday and Hasan 1985).

Conjunctive relations can be encoded by either structural resources of interdependency within a sentence (through paratactic and hypotactic linking of clauses) or by non-structural resources of cohesion between sentences (through a conjunctive adjunct or certain conjunctions like *and*, *or*, *but* etc.) and realized in three main modes of expression: elaboration (+) (restating/clarifying), extension (=) (adding/varying) and enhancement (x) (expanding in time, reason, condition, concession) (Bloor and Bloor 1995; Eggins 2000; Halliday 1985).

Since the child reader's perception of cohesive ties improves developmentally with age (especially btw 8-13) (Chapman 1982), it is essential to determine the kind of cohesive devices preferred in school texts, and also to reveal the nature of conjunctive relations in textbooks for different grades. For this reason, the purpose of this study was to find out: (1) how conjunctive relations were developed in discourse, (2) what kind of conjunctive relations were chosen, and (3) whether lexicogrammatical choices changed among similar text types and in different grades by comparing extracts from Turkish and English pedagogical texts.

2 Method

The descriptive analysis method was adopted in this study for investigating patterns of conjunctive relations. The corpus consisted of narrative, biographical and informative texts, derived from the three units of two coursebook series for 11-14 year-olds: the locally-produced, *Ilkogretim Turkce 6-7-8*, and the world-renowned, *Cambridge Checkpoint English 7-8-9*, in native-language instruction (Cox

2012, 2013, 2014; Deniz 2013a, 2013b). After the extracts were converted to Word and divided into sentences, conjunctives were given functional labels, and frequencies and percentages were calculated with Excel.

3 Results of the conjunctive analyses

When the realization of conjunctive relationships was examined in the narrative extracts from the sixth-graders' textbooks, it was found as in Table 1 that the Turkish narrative made more use of cohesion (56.55%) than interdependency (49.18%) both within and across categories (3.44%).

Rather than building up conjunctive relations within clause complexes, the Turkish narrative employed non-structural resources between clause complexes, and the dominant categories were extension (30.32%) and enhancement (24.59%) in the case of cohesion. Likewise, interdependency was much preferred for building conjunctive relations of extending (10.65%) and enhancing types (38.52%) in the Turkish narrative.

Narrative	COH		Exemplars	INTERD		Exemplars
TR	f	%	TR	f	%	TR
+	2	1.63	daha doğrusu (1), üstelik (1)	0	0	-
=	37	30.32	bunun yerine (1), oysa (2)	13	10.65	ama (4), ve (3), kaç-Ip (5)
x	30	24.59	bundan sonra (1), böylece (1)	47	38.52	çünkü (2), için (10), eğer (3), sanki (1)
∑ items	69	56.55		60	49.18	
∑sentences	122	100		122	100	
EN	f	%	EN	f	%	EN
+	0	0	-	1	1.72	which (1)
=	1	1.72	but (1)	15	25.86	and (13), but (1), not only but so (1)
x	1	1.72	otherwise (1)	18	31.03	when (8), although (1), unless (1)
∑ items	2	3.44		34	58.62	
∑sentences	58	100		58	100	

Table1: Conjunctive relations in narrative texts

Unlike the Turkish, the English narratives used cohesive items only on two occasions (3.44%), and focused more on the formation of interdependencies between clauses (58.62%). Although elaboration was either rare or even non-existent in both samples, extension and enhancement dominated not only in the Turkish but also in the English narratives, especially if in the form of interdependent clauses (25.86%, 31.03%). It is evident from Table 1 that there was a more balanced distribution of cohesive and interdependent items in the Turkish, compared to the English narratives, where interdependencies dominated. In other words, the Turkish text availed all the opportunities to create conjunctive relations across larger parts of the text without recourse to structural resources, whereas the English texts paid particular attention to forming conjunctive relationships within the boundaries of the clause complex itself, making it easier to track semantic links. Also, the lack of elaboration in each situation can be attributed to the fact that narratives were more concerned with giving information about the details of events rather than clarifying or restating facts or statements.

Table 2 displays the results of the conjunctive analyses from biographical texts in seventh-graders' textbooks. A comparison of the two corpora with respect to the use of cohesion indicated that the Turkish biographical text (%32.60) surpassed the English ones (12.72%),.

Biographical	COH		Exemplars	INTERD		Exemplars
TR	f	%	TR	f	%	TR
+	1	2.17	bunun yanı sıra (1)	0	0	-
=	7	15.21	da (6), bile (1)	10	21.73	ve (3), ancak (1), yüksel-ErEk (5)
X	7	15.21	daha sonra (3), bunun üzerine (1)	6	13.04	için (3), iken (2), -mEk üzere (1)
∑ items	15	32.60		16	34.78	
∑sentences	46	100		46	100	
EN	f	%	EN	f	%	EN
+	0	0	-	9	16.36	where (4), consider-ed (5)
=	3	5.45	but (1), however (2)	22	40	and (13), but (6), or (2)
X	4	7.27	then (1), finally (1)	20	36.36	after (3), despite (2), if (1)
∑ items	7	12.72		51	92.72	
∑sentences	55	100		55	100	

Table 2 Conjunctive relations in biographical texts

while they were both deficient in the use of elaboration, and had a liking for extending (15.21%, 5.45%) and enhancing (15.21%, 7.27%) types of cohesive resources

As for the comparison of the Turkish and English biographical texts with regard to the use of interdependent clauses, the English sample (92.72%) outnumbered the Turkish one (34.78%). It is clear from Table 2 that extending (21.73%, 40%) and enhancing (13.04%, 36.36%) categories of interdependency took the lead, as opposed to elaborating type of interdependency (0%, 16.36%), appearing seldom in both kinds of extracts again.

As in the case of the Turkish narrative, there is an even distribution of non-structural and structural resources of conjunction in the Turkish biographical text. However, a substantial increase was observed in the amount of cohesive and interdependent items in the English biographical texts.

Informative	COH		Exemplars	INTERD		Exemplars
TR	f	%	TR	f	%	TR
+	5	6.49	söz gelimi (3), hatta (2)	0	0	-
=	19	24.67	ne var ki (3), benzer biçimde (1)	20	25.97	olsun...olsun (1), -mAk yerine (1)
x	22	28.57	dolayısıyla (1), başta (2), sonra (1)	21	27.27	için (8), -Ir gibi (1), -sE dE (4)
∑ items	46	59.74		41	53.24	
∑sentences	77	100		77	100	
EN	f	%	EN	f	%	EN
+	0	0	-	8	17.39	which (1), resembl-ing (7)
=	6	13.04	actually (1), also (1), though (1)	11	23.91	and (9), instead of (1)
x	7	15.21	therefore (2), then (1)	16	34.78	to save (8), only when (1), so (2)
∑ items	13	28.26		35	76.08	
∑sentences	46	100		46	100	

Table3: Conjunctive relations in informative texts

This might be related to the choice of life stories told. In Mehmet Akif's (the poet of the Turkish national anthem) bio, his career details were extended in chronological order and elements of his literary style were exposed to the reader, whereas in the English corpus, opposition leaders like Nelson Mandela (South Africa's first black president) and Aung San Suu Kyi (Myanmar's Nobel Peace Prize winner) were introduced with frequent references to their struggle against dominant forces, possibly requiring the construction of a more explanatory text in regard to causes and consequences.

Table 3 summarizes findings from the conjunctive analysis of informative texts. The distributive pattern of cohesion and interdependency in biographical texts was preserved in informative texts: a higher concentration of cohesion in the Turkish sample (59.74%, 28.26%), and a denser population of interdependency in the English version (76.08%, 53.24%) along with almost equal uses of structural (f=41) and non-structural (f=46) resources of conjunction in the Turkish informative texts.

As can be seen from Table 3, elaboration was the least frequently-used category, whether it was realized structurally (17.39%) or non-structurally (6.49%) in both kinds of informative texts as usual.

4 Conclusion

The conjunctive analyses of the pedagogical texts extracted from Turkish and English coursebooks for native-language instruction revealed that: (1) the English texts had the tendency to create conjunctive relationships by forming both structural and semantic links between clauses and avoided breaching sentence borders by using interdependency resources predominantly; and (2) the Turkish texts were, however, inclined to capitalize on structural and non-structural resources non-sparingly, and the construction of conjunctive relations merely on the semantic level might lead to comprehension difficulties in the inexperienced child reader, who would need to look for more linguistic clues to hold onto while struggling to process meaning relationships within the larger domain of discourse beyond the sentence level.

Moreover, regardless of the text type, elaborating mode of conjunctive realizations were seen rare in both languages, as the generic characterization of the texts was more oriented towards developing topic in terms of time, manner, reason and adversity, instead of reinstating. Despite the absence of such particularity in the Turkish extracts, the English pedagogical texts tended to gradually make more use of cohesion as grade level improved. Consequently, it would be advisable for the Turkish materials writer to put conjunctive choices on a more systematic basis because young readers' text comprehension can be enhanced if their decoding skills are supported by high cohesion (McNamara et al. 2011). Since increased cohesion facilitates comprehension and recall, naive readers could tackle processing problems if explicit textual clues, connectives, were made available to them, especially in the face of difficult texts (O'Reilly and

McNamara 2007).

References

Bloor, T. and Bloor, M. 1995. *The functional analysis of English*. New York: Oxford University Press Inc.

Chapman, L. J. 1982. "A study in reading development: A comparison of the ability of 8, 10 and 13 year old children to perceive cohesion in their school texts". *The Annual Meeting of the United Kingdom Reading Association*, 19-23.07.1982, Newcastle upon Tyne, England.

Cox, M. 2012. *Cambridge checkpoint English coursebook 7*. Cambridge: Cambridge University Press.

Cox, M. 2013. *Cambridge checkpoint English coursebook 8*. Cambridge: Cambridge University Press.

Cox, M. 2014. *Cambridge checkpoint English coursebook 9*. Cambridge: Cambridge University Press.

Deniz, K. (Ed.) 2013a. *İlköğretim Türkçe ders kitabı 6*. Ankara: MEB Devlet Kitapları.

Deniz, K. (Ed.) 2013b. *İlköğretim Türkçe ders kitabı 7*. Ankara: MEB Devlet Kitapları.

Eggins, S. 2000. *An introduction to systemic functional linguistics.* New York: Continuum.

Gamble, N. and Yates, S. 2002. *Exploring children's literature: Teaching the language and reading of fiction*. London: Paul Chapman Publishing.

Halliday, M. A. K. (1985). *An introduction to functional grammar*. London: Edward Arnold.

Halliday, M. A. K. and Hasan, R. 1985. *Cohesion in English.* New York: Longman Inc.

Halliday, M. A. K. and Matthiessen, C. M. I. M. 2004. *An introduction to functional grammar (Third edition).* London: Arnold.

McNamara, D. S., Ozuru, Y. and Floyd, R. G. 2011. "Comprehension challenges in the fourth grade: The roles of text cohesion, text genre, and readers' prior knowledge". *International Electronic Journal of Elementary Education,* 4(1), 229-257.

O'Reilly, T. and McNamara, D. S. 2007. "Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers". *Discourse Processes: A Multidisciplinary Journal*, 43(2), 121-152.

Şahin, D. 2013. *İlköğretim Türkçe ders kitabı 8*. Ankara: Ada Matbaacılık.

A corpus-based discourse analysis of representations of mental illness and mental health in the British Press

Gillian Smith
Lancaster University
g.smith6@lancaster.ac.uk

1 Introduction

A topic of recent interest has been the stigmatisation of mental illness. The British press have been accused of perpetuating this, providing the public with negative representations of mental illness based on misguided stereotypes (Bilić and Georgaca, 2007; Nawková et al., 2001; Stuart, 2003; Thornton and Wahl, 1996; Coverdale et al., 2002). This paper presents a corpus-based analysis of representations of mental health in the British press between 2011 and 2014, aiming to broaden the scope of earlier works, using a larger, more representative sample and a discourse approach.

2 Literature Review

The discourses surrounding mental illness are interesting, with the World Health Organisation (2005:5) claiming that 'people who experience them [mental health problems] still meet fear and prejudice'. This stigma arguably stems from society's construction of discourses surrounding mental illness, a major source of which is the media (Nairn et al.,2001; Hallam, 2002; Olstead, 2002; Anderson, 2003; Nawková et al., 2001). This is problematic, as studies have shown that media portrayals of mental illness tend to be considerably negative (Nawková et al., 2001; WHO, 2005; Bilić and Georgaca, 2007; Nawková et al., 2001; Stuart, 2003; Thornton and Wahl, 1996; Coverdale et al., 2002). Such negative portrayals, Nawková et al. (2001:23) note, lead to a 'distorted picture' of mental illness, which in turn inspires discrimination. Some, however, have acknowledged that the media may be a source for destigmatisation (Byrne, 2000; Stuart, 2003; Anderson, 2001; Bilić and Georgaca, 2007), through raising awareness of discrimination and therefore acting as a catalysts of change in the representations of mental illness.

3 Methodology

This project involved the collection of a British newspaper corpus. Texts were collected from the online news archive *Nexis: News Search* (2014) using the search term: mental illness* OR mental health* OR bipolar OR schizophreni* OR manic depression OR cognitive behavioural OR cbt OR

postnatal depression OR post natal depression OR pnd OR posttraumatic stress OR post traumatic stress OR ptsd OR obsessive compulsive disorder* OR obsessive compulsive OR ocd OR personality disorder* OR seasonal affective disorder* OR self harm* OR bulimia OR agoraphobi* OR psychiatr*
All articles containing these queries published between 1st January 2011 and 31st December 2013 were collated from each of the UK's six highest circulating newspapers– The Sun, The Mail, The Mirror, The Guardian, The Telegraph and The Times – amounting in a corpus of just over eighteen million words.

This corpus-assisted discourse analysis aims to 'analyze discourse practices that reflect or construct social problems' (Bloor and Bloor, 2007:12). Hence, corpus methods were used to reveal dominant hegemonic discourses in press reports of mental illness. Collocations analysis revealed meanings surrounding the terms 'mental illness' and 'mental health' within the corpus, with collocations grouped in terms of their semantic preference, allowing identification of the semantic prosody of the term, which in turn highlights the discourses surrounding mental illness.

4 Results

Grouping collocates of 'mental illness' and 'mental health' revealed five central discourses surrounding the term:

- problematizing the extent of mental illness
- problematizing the effects of mental illness
- medicalization
- stigma and discrimination
- awareness

The first discourse shows the press problematizing the extent of mental illness, focusing upon notions of severity, both in terms of prevalence and impact. Firstly, this discourse includes evaluations of the severity of mental illness, highlighted through concordance analysis of the collocates 'severe' and 'serious'. In all instances, the terms are used as premodifiers, suggesting that the press focus on the more extreme forms of mental illness. Furthermore, they are often used as marker of real or non-existent problems, as in examples like 'no symptoms indicative of serious mental illness' or 'did not meet the criteria of serious mental illness'. This ignores the spectrum of mental illnesses in favour of extreme forms.

Secondly, within this discourse, the press demonstrate a tendency to focus upon the extreme impact of mental health problems. Both 'struggle' and 'suffer' are collocates. These verbs render the mentally ill passive victims of an aggressive illness, as semantically they concern some form of physical effort. Thus, mental illnesses are surrounded by a negative discourse that portrays them as both severe in nature and as life limiting.

Another discourse that portrays mental illness as problematic involves press discussion of social issues that tend to co-occur with mental illness. The terms 'violent' and 'violence' are often compared to mental health through conjunctions and listing structures, such as 'violent behaviour and mental illness' or 'a history of mental illness, violence, alcohol abuse'. This is problematic, as Thornton and Wahl (1996:17) note that press' construction of mental illness and violence as comparable consequently inspires fear'. The undesirable collocates 'substance [abuse]', 'alcoholism' and 'addiction' were also frequently linked to mental health using syndetic and asyndetic listing, such as 'mental illness, drug addiction and alcoholism'. This habitual co-occurrence in listing structures leads to a comparison and mapping of characteristics between mental illness and these problems, which leads to mental illness being viewed as socially deviant in the same ways addiction is, perceived as self-inflicted, excessive and harmful.

Another discourse surrounding mental health revealed by its collocates' semantic preferences and mentioned extensively in the literature is medicalization. This discourse involves the use of medical terminology to describe non-biological experiences, which, as Bilić and Georgaca (2007:167) note, 'constructs "mental illness" as a medical disorder, psychiatrists as responsible for its management, and people with mental health problems as passive sufferers of their condition'.

Two thirds of collocates within the medicalization discourse fall into the symptomatology subsection, where 'signs' and 'symptoms' are collocates of mental illness/health. Signs/symptoms of mental illness in this sample are again an indicator of being mentally ill or not, such as 'no history or symptoms of mental illness' or 'showed no signs of mental illness'. Again, this ignores the spectrum of mental health problems, implying mental illness is one disease with a single set of symptoms and only those with these symptoms are ill.

The second subsection of the medicalization discourse is medical treatment. The term 'treating' usually concerns mental health treatment as something which 'costs the NHS' and 'may be expensive', portraying mental illness as a costly burden. The noun 'treatment' is used to discuss those who 'specialise' in mental health treatment, suggesting mental illness is complex, requiring professional treatment, which in turn serves to isolate mental health as different. This creates a medicalized discourse surrounding the terms, where symptoms are taken as indicators of 'real' illness

and treatment is discussed as being costly and specialist, generating a negative discourse surrounding mental illness as burdensome.

Another discourse of mental illness/health is stigma and discrimination, where collocates may be semantically grouped to reveal stereotypes deemed representative of public opinion, which specifically concern prejudiced attitudes. Two of the most significant collocates are 'stigma' and 'discrimination', which were identified as prominent discourses within the literature. Interestingly, however, when we look through a selection of concordance lines, these do not promote stigma, but rather aim to highlight and abolish it. Stigma is often discussed alongside terms which call for its demise, such as 'break down the stigma', end the stigma', 'reduce the stigma' and 'tackling stigma and discrimination'. This highlights that the media are not only raising awareness through discussing and therefore exposing mental health discrimination, but also actively campaign for its eradication, rather than perpetuating discrimination.

The final discourse identifiable from semantic preferences of mental illness/health collocates within the corpus is awareness, where the press aim to improve understanding both of mental illness in general and of charities concerned with mental health. As identified in the literature (Byrne, 2000; Stuart, 2003; Anderson, 2001; Harper, 2009; Coverdale et al., 2002; Bilić and Georgaca, 2007), the media, whilst being accused of perpetuating negative stereotypes of mental illness, have also been acknowledged as a potential source for change in representations. Firstly, collocates of mental illness reveal the two major UK mental illness charities, Mind and Rethink, which the press cite as sources of research and as offering 'advice' and 'information', as well as noting they 'campaign' against stigmatization. More general collocates concern raising awareness both of mental illness and the stigma surrounding it. Concordance lines of 'understanding' and 'awareness' both reveal calls for an increase in knowledge of mental health, using terms of growth like 'increases', 'breakthrough', 'wider', 'raise', 'raising' and 'better'. This importantly highlights the press both acknowledging the need for and demanding more awareness of mental illness, despite the negative portrayals discussed earlier.

5 Conclusion

These findings highlight that press representations of mental illness are considerably negative, which in turn perpetuates the stigma surrounding mental illness, as the press' misrepresentations are the predominant source of public information. However, the stigma and awareness discourses concern press discussion of the need for wider understanding and suggests that, whilst the press portray mental illness in discriminatory ways, they attempt to change public opinion. It may be suggested, however, that for the press to raise full awareness, they first must address their own stigmatizing representations.

References

Bilić, B & Georgaca, E. (2007). Representations of "Mental Illness" in Serbian Newspapers: A Critical Discourse Analysis. *Qualitative Research in Psychology, 4(1-2)*, 167-186.

Bloor, M. & Bloor, T. (2007). *The Practice of Critical Discourse Analysis.* London: Hodder Education.

Byrne, P. (2000). Stigma of mental illness and ways of diminishing it. *Advances in Psychiatric Treatment*, *6,* 65-72.

Coverdale, J., Nairn, R., and Claasen, D. (2002). Depictions of mental illness in print media: a prospective national sample. *Australian and New Zealand Journal of Psychiatry, 36,* 697–700.

Hallam, A. (2002). Media influences on mental health policy: long-term effects of the Clunis and Silcock cases. *International Review of Psychiatry, 14,* 26–33.

Harper, S. (2009). *Madness, power and the media: class, gender and race in popular representations of mental distress.* Basingstoke: Palgrave.

Nairn, R., Coverdale, J., & Claasen, D. (2001). From source material to news story in New Zealand print media: a prospective study of the stigmatizing processes in depicting mental illness. *Australian and New Zealand Journal of Psychiatry, 35,* 654–659.

Nawková, L., Nawka, A., Adámková, T., Rukavina, T.V., Holcnerová, P., Kuzman, M.R…. Raboch, J. (2001). The picture of mental health/illness in the printed media in three Central European countries. *Journal of Health Communication*, *17(1),* 22-40.

Nexis: News Search. (2014). Retrieved from http://www.lexisnexis.com/uk/nexis/search/loadForm.do?formID=GB01NBSimplSrch&random0.7450459500123497. Date accessed: 20th October 2014.

Olstead, R. (2002). Contesting the text: Canadian media depictions of the conflation of mental illness and criminality. *Sociology of Health and Illness, 24,* 621–643.

Stuart, H. (2003). Stigma and daily news: evaluation of a newspaper intervention. *Canadian Journal of Psychiatry, 48,* 651–656.

Thornton, J.A. & Wahl, O.F. (1996). Impact of a newspaper article on attitudes toward mental illness. *Journal of Community Psychology, 24,* 17–25.

World Health Organisation (2005). *Mental Health – Facing the challenges, building solutions.* Geneva: World Health Organisation.

A Multi-Dimensional Comparison of Oral Proficiency Interviews to Conversation, Academic and Professional Spoken Registers

Shelley Staples
Purdue University
slstaples@purdue.edu

Jesse Egbert
Brigham Young University
Jesse_Egbert@byu.edu

Geoffrey T. LaFlair
University of Kentucky
gtl7@nau.edu

1 Introduction

The use of oral proficiency interviews (OPIs) to measure speaking ability has been rationalized by the argument that they mirror aspects of interactive spoken discourse (see e.g., Kasper and Ross, 2007). However, discourse analysts have, since the 1990s, provided evidence that OPIs constitute their own speech event, weakening inferences about test taker abilities to the domain of conversation (e.g., van Lier, 1989). However, these studies have mostly focused on qualitative analysis. In addition, extensive research has shown language to be multi-componential in nature and that it is often best described in terms of co-occurrence patterns among multiple linguistic features (Biber, 1988, 2006). No quantitative analyses of OPIs that consider constellations of linguistic features have been conducted in relation to other spoken registers.

Multi-Dimensional (MD) analysis is a methodological approach that relies on factor analysis to identify functionally interpretable factors (dimensions) of linguistic variation to compare registers. Biber's (1988) first dimension contrasted the co-occurrence of linguistic features such as pronouns and adverbials in speech with the co-occurrence of features such as nouns and attributive adjectives in writing. Importantly, these features serve very different functions, on the one hand indicating involvement in an interaction and on the other providing detailed information to readers. While many MD analyses have emphasized differences across speech and writing (Biber, 1988, 2006), recent studies have also revealed important differences in spoken registers (Friginal, 2009; Al Surmi, 2012). This research indicates that there is a great deal of variation within spoken interactive discourse yet to be examined.

In this study we use MD analysis to investigate a large number of lexico-grammatical features used in one OPI, the Michigan English Language Assessment Battery (MELAB) speaking assessment. The study aims to identify linguistic and functional characteristics of the MELAB in relation to conversation as well as other interactive spoken registers that reflect the academic and professional purposes (e.g., nursing) for taking the MELAB OPI. The findings have implications for the study of register variation across spoken interactive discourse as well as for language assessment.

2 Methods

An MD analysis was conducted to compare a sample of 98 MELAB OPIs with (a) conversation (N = 716); (b) nurse-patient interactions (N = 50); (c) office hours (N = 11); (d) study groups (N = 25); (e) service encounters (N = 22). The MELAB corpus was developed in 2014 by the authors in coordination with Cambridge Michigan Language Assessment (CaMLA). The conversation corpus comprises the Longman Corpus of American Conversation (Biber et al., 1999). The American Nurse-Standardized Patient (ANSP) corpus was developed by one of the authors in 2012. The office hours, study groups, and service encounters consist of three sub-corpora from the T2K-SWAL corpus (Biber, 2006). For the present study, the MELAB-OPI and ANSP corpora were also divided by speaker group.

Linguistic features chosen for the MD analysis were identified based on previous MD analyses and a pilot study of individual features across the four registers (LaFlair, Egbert, and Staples, 2014). The final analysis included pronouns, contractions, stance devices (e.g., adverbials and modals), interactional features (e.g., questions), nouns, nominalizations, attributive adjectives, noun + OF phrases (e.g., source of water), and relative clauses.

After the individual linguistic features were chosen, factor analysis was performed on the normed rates of occurrence. Using the statistical software R, we conducted the factor analysis using principal axis factoring and a Promax rotation. The scree plot of eigenvalues revealed a definitive break between the sixth and seventh factor. Therefore, a six factor solution was chosen. The cumulative variance accounted for by the six factors was 43%. Variables were only included in the analysis if they met a minimal factor loading threshold of +/-.30. After assigning each of the variables to the factor in which it loaded the strongest, the positive-loading features were separated from the negative-loading features.

After establishing the factor structure, we calculated dimension scores for each of the texts in the four corpora, using standardized rates of occurrence for each linguistic feature. The final step in the MD analysis was to explore the underlying functional interpretation of each factor and assign a dimension

label to each of the factors. We relied on two sources of information to complete this step for each of the six dimensions: (a) previous research on the co-occurring features that loaded on the dimension, and (b) the use of these linguistic features in the texts.

3 Results and Conclusions

We identified six dimensions from the MD analysis of the 4 corpora:

- (1) Explicit expressions of stance;
- (2) Future possibilities;
- (3) Speaker centered informational discourse vs. Listener centered involvement;
- (4) Extended informational discourse;
- (5) Expression of personal desires vs. Narratives
- (6) Implicit expressions of stance.

We will only examine Dimension 1 here. Expressions of stance (e.g., adverbials, modals, and complement clauses) have long been identified as characteristic of conversation (Biber, 1988). Speakers can express their stance (personal feelings, attitudes, and evaluations) more overtly/explicitly by using features such as first person pronouns and stance verbs (e.g., *I think that*), that overtly mark the agent (the speaker). Alternatively, speakers can express stance more implicitly, by using adverbials (*certainly, actually*), in which the speaker does not need to be overtly identified as the agent of the stance (Biber et al., 1999, p. 864-865). The features loading onto Dimension 1 included all verb + *that* complement clauses, as well as *that* clauses associated with specific semantic classes of verbs (e.g., certainty verbs such as *know*). In addition, mental verbs (e.g., *think*) loaded onto this dimension. Finally, *that* deletion was also strongly associated with Dimension 1. As Figure 1 shows, the highest dimension scores for Dimension 1 were found for patients within the ANSP corpus (1.74) and conversation from the Longman corpus (1.41).
Patient speech and conversation was characterized by a greater use of explicit stance:

(1) Nurse: Well there's a history of heart problems in your family?
Patient: I *know (that) my mom and I have hypertension*. [ANSP corpus]
(2) Speaker A: Yeah, yeah. I *guess (that) it just makes people kind of nervous* just
Speaker B: I *think what's xxx that it's um*. I *guess (that) for me, I'm so conscious of my speaking voice* and how.
Speaker A: Oh really?
[Longman conversation corpus]
Both patients and speakers in face-to-face conversation use explicit stance to clearly identify the agent (themselves, in these two cases). However, the function of these stance features are quite different across the two speaker groups. For patients, it is important to explicitly identify themselves as the source of knowledge for the nurse. In conversation, explicit stance may be used to build rapport with the other speaker, and to show shared feelings and attitudes, which are explicitly identified as coming from the speaker.

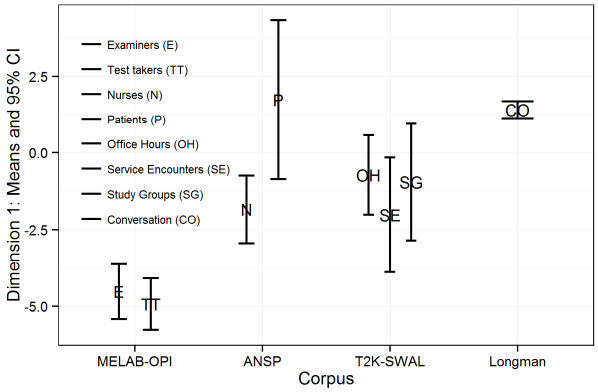


Figure 1. Dimension 1 scores across the four corpora.

In contrast, both test takers (-4.93) and examiners (-4.52) in the MELAB corpus use less explicit stance than all of the other registers. This is reflected in the following example, where there are no instances of *that* complement clauses and only one mental verb used (*see*):

> T: With the geriatric patients you go more for the maintenance and keeping them in comfort
> E: Uh huh.
> T: But in clinics you get the like definite results more improvement
> E: Right. Yes.
> T: So it motivates you to work more in my opinion.
> E: Right? Okay yes absolutely yeah yeah yeah so so you actually *see* full recovery at clinics
> T: yes
> E: where you won't with geriatrics
> T: yeah.

The use of less explicit stance in the MELAB is likely due to the different purposes of OPIs. Rapport building would certainly be less of a concern in an OPI, as there is no reason to assume that the examiner and test taker will ever meet again. In addition, although information sharing is a clear function of the OPI, the source of the information may be less important. In fact, it may behoove test takers (who scored lowest on this dimension) to express stance less explicitly if at all. This is supported by the findings of LaFlair, Staples, and Egbert (2015) that higher scoring test takers on the MELAB used less explicit stance.

4 Conclusion

The findings indicate that the MELAB OPI has distinct characteristics in relation to other spoken interactive discourse. These differences have implications for the extrapolation inferences for the MELAB and other OPIs (Kane, 2013). In addition, the results show that there is a great deal of variation across the spoken interactive registers investigated in this study. The quantitative findings provide important insight into the linguistic features and the functions they represent. Additional qualitative analysis of individual texts is necessary to more fully understand this variation and its relation to the situational characteristics of each register and speaker group. Along with qualitative analysis, future research should explore the differences in other speaker groups (e.g., professors and students in office hours). Finally, over half of the variance in the registers and speaker groups could not be explained by the MD analysis presented here. This indicates that other linguistic features (e.g., speech rate or prosody) may play a role in differentiating the registers.

Acknowledgements

This research was partially funded by a Cambridge Michigan Language Assessment SPAAN grant.

References

Al Surmi, M. 2012. Authenticity and TV shows: A multidimensional analysis perspective. *TESOL Quarterly*, 46 (4), 671-694.

Biber, D. 1988. *Variation across speech and writing*. Cambridge: CUP.

Biber, D. 2006. *University language: A corpus-based study of spoken and written registers*. Philiadelphia, PA: John Benjamins Publishing.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. 1999. *The Longman grammar of spoken and written English*. London: Pearson.

Friginal, E. 2009. *The language of outsourced call centers: a corpus-based study of cross-cultural interaction*. Amsterdam: John Benjamins.

Kane, M.T. 2013. Validating the interpretations and uses of test scores. *Journal of Educational Measurement,* 50 (1), 1-73.

Kasper, G. and Ross, S.J. 2007. Multiple questions in oral proficiency interviews. *Journal of Pragmatics*, 39, 2045-2070.

LaFlair, G., Egbert, J., and Staples, S. 2014, September. *Comparing oral proficiency interviews to academic and professional spoken registers*. Paper presented at the meeting of AACL, Flagstaff, AZ.

LaFlair, G., Staples, S. and Egbert, J. 2015. *Variability in the MELAB speaking task: Investigating linguistic characteristics of test taker performances in relation to rater severity and score*. CaMLA Working Papers.

van Lier, L. 1989. Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*, 23(3), 489-508.

"Do you like him?" "I don't dislike him." Stance expression and hedging strategies in female characters of Downton Abbey. A case study

Anna Stermieri
University of Modena e Reggio Emilia
anna.stermieri@unimore.it

Cecilia Lazzeretti
University of Modena e Reggio Emilia
cecilia.lazzeretti@unimore.it

This work aims at investigating the ways in which characters are constructed in dialogue in *Downton Abbey*, a British period drama television series featuring the lives of an aristocratic family and of their servants in the Yorkshire of the early 20th century. The background for the study includes Bednarek studies on televisual characters characterization (Bednarek 2011, 2012) and Culpeper seminal work on language and fictional characters characterization (2001).

The investigation focusses on the characterization of two female characters which are opposite in terms of intended social position and personality, but which have been chosen because of their centrality to the show (they are both permanent characters, appearing in all the episodes of Season 1 and 2, and are still part of the story and of the cast). It is important to keep in mind that the presence of Character 1 is much heavier than Character 2 as demonstrated by the number of words spoken by each of them (Character 1: 16.162 words vs. Character 2: 4.466). It is interesting however to compare them as they represent two opposite yet complementary aspects of femininity in the society the show aims to recreate. In this analysis we thus set out to explore the aspects of femininity which are displayed by means of the rhetorical and lexical choices guiding the speech of such characters.

The study is based on a small corpus of 226.490 words, including the lines of the two characters for the first three series of the show. Drawing upon a combined methodology, based on both qualitative discourse analysis and quantitative corpus methodologies, in line with the Corpus Assisted Discourse Studies tradition (Partington et al. 2013), we aim at unveiling the set of values behind the characters investigated, by uncovering the linguistic cues that reveal their personality. In addition, we also aim at shedding light upon the discoursal structures which are exploited in the construction of the characters and in their positioning in the network of relationships governing the narrative on screen.

Preliminary results show a tendency for both characters to rely on interrogative sentences and negative sentences.

For example Character 1 asks a question every five sentences, while Character 2 asks a question every four sentences. Also interesting is the use of question tags, (8% of the questions of Character 2 and 3% of the questions of Character 1). This aspect might represent a linguistic trait of Britishness in movies (see Chiaro 2000: 30).

In addition, Character 1 has shown a widespread use of negative statements with 18% of her statements beginning with *I* in the form: *I* + *aux/modal* + *negative* + *predicate*, interestingly associated with expressions of thought, feeling, opinion and will, such as *know* (18 occurrences); *think* (17 occurrences); *sure* (10 occurrences); *want* (10 occurrences).

This seems to suggest a character construction heavily relying on indirectness and hedging. For example, one of the most frequent word clusters for this character is *I am afraid* (17 occurrences). This aspect seems in line with Lakoff's argument that women have a stronger tendency than men towards hedging thus preferring strategies aimed at avoiding strong statements (1975).

Character 2 displays an evident tendency towards the use of negative constructions as well. The analysis showed that 24% of the occurrences of *I* fulfil the pattern *I* + *aux/modal* + *negative* + *predicate*. In addition, 5 different negatives appear among the top hundred words in the frequency list for Character 2 (e.g. *'t,* occurring 129 times or *not* occurring 47 times), attesting for 5% of her overall spoken words.

This widespread use of negation on the part of Character 2 might reflect what Weintraub has pointed out, i.e. that "[…]speakers who use many negatives tend to be oppositional and stubborn" (2003: 145) and that the expression of anger can be associated with a high use of negatives (see Weintraub 1989). This description seems to fit Character 2, who is not completely happy with her job and is less involved in the plot. Moreover, it seems to reveal her feelings of dissatisfaction and insecurity, both at work and in her love life.

The tendency of Character 2 towards a less hedged expression of stance, in comparison to the extensive use of hedging in Character 1, might be linked to the different social status of the two women. Character 1 is in fact a member of the aristocracy, whereas Character 2 is one of the servants living in the estate.

The representation of the two women seems to be influenced by the different expectations that society (in fiction) and audience (in real life) have towards them. Moreover, the series is a historical drama aimed at a modern audience, which might not be

entirely familiar with the social conventions of the period. The audience may be therefore engaged by aspects of nostalgia and cultural belonging (Baena and Byker 2014).

References

Baena, R. and Byker, C. 2014. “Dialects of nostalgia: Downton Abbey and English identity”. National Identities, (ahead of print) 1-11.

Bednarek, M. 2012. "Constructing Nerdiness: Characterisation in "The Big Bang Theory". Multilingua: Journal of Cross-Cultural and Interlanguage Communication 31 (2): 199-229.

Bednarek, M. 2011. "Expressivity and televisual characterization". Language and Literature 20(1): 3-21.

Culpeper, J. 2001. Language and characterisation: people in plays and other texts. London: Longman.

Chiaro, D. 2000. "The British will use tag questions, won’t they? The case of Four Weddings and a Funeral". Tradurre il Cinema. Trieste: Università degli Studi di Trieste: 27-39.

Lakoff R. 1975. Language and Women's Place. New York: Harper Row.

Partington, A., Duguid, A. and Taylor, C. 2013. Patterns and meanings in discourse. Theory and Practice in Corpus-Assisted Discourse Studies (CADS). Amsterdam: John Benjamins.

Weintraub, W. 1989. Verbal Behavior in Everyday Life. New York: Springer.

Weintraub, W. 2003. “Verbal Behaviour and Personality Assessment.” In Post, J.M. (ed) The Psychological Assessment of Political leaders, The University of Michigan Press: University of Michigan: 137-153.

An initial investigation of semantic prosody in Thai

Pornthip Supanfai
Lancaster University
p.supanfai1@lancaster.ac.uk

1 Introduction

To date, there have been relatively few studies of semantic prosody in languages other than English, especially in Asian languages. This paper explores the nature of semantic prosody in Thai, using two contrasting approaches. The aim is (a) to arrive at an understanding of the semantic prosodies of a number of lexical units; (b) to assess what approach to this phenomenon produces better results in the study of semantic prosody in Thai.

2 Background: the two approaches

Semantic prosody has become an important concept in corpus linguistics (Whitsitt 2005: 283; Bednarek 2008: 119), and it has attracted much interest in the past 15 to 20 years (Ebeling 2014: 161; Stewart 2010: 6). Because it is a relatively new concept, there is no consensus on its definition (Zhang 2009: 2), and even its name is controversial (Partington 2014: 279). Primarily, four scholars - namely Louw (1993), Sinclair (2004), Stubbs (1995; 2001), and Partington (1998; 2004; 2014) - have substantially contributed to discussion of the concept.

Louw was the first person to introduce the term *semantic prosody* to the public, although he credits Sinclair for having provided him with the term and the concept (Whitsitt 2005: 283-186). Louw (1993: 159) defines semantic prosody as “a consistent aura of meaning with which a form is imbued by its collocates”. In particular, a lexical item may be said to display either a *positive* or *negative* semantic prosody, depending on the context it habitually occurs in. However, Louw’s definition of semantic prosody is different from those of Sinclair, Stubbs, and Partington, who also differ from one another.

Most scholars, including Louw, Stubbs, and Partington, assert that they follow Sinclair’s approach to semantic prosody. It can, however, be argued that in reality they do not completely adopt Sinclair’s approach. Louw, Partington and Stubbs tend to identify semantic prosody by analysis of individual co-occurring words (“collocates”) and restrict semantic prosody to being positive or negative. Sinclair, on the other hand, identifies semantic prosody from pragmatic meaning(s) which may be spread across an extended unit of meaning; moreover he does not confine the pragmatic meaning that is conveyed to the positive vs. negative

opposition (see, for instance, Sinclair 2004: 34). It may thus be argued that there exist two contrasting approaches to semantic prosody in the literature, one represented by the work of Louw, Stubbs, and Partington, and the other by the work of Sinclair and also endorsed by Hunston (2007: 257).

3 Method

The corpus used in this study was the Thai National Corpus. This is a general corpus that represents present-day standard written Thai and consists of around 33 million words (Aroonmanakun 2007: 4).

In order to test the applicability of each of the two approaches outlined above, I looked at a small sample of three word types: /kreeŋcay/ ('(be) considerate'), /kɔ̀ɔhâykɛ̀ɛt/ ('cause') and /chɔ̂ɔp/ ('like'). Each of these words was selected on the basis of a different motivation. The word /kreeŋcay/ is interesting because there seems to be no word in English that has exactly the same meaning as /kreeŋcay/. The closest translation equivalent would probably 'considerate' or 'reluctant (to impose on a person)'. /kɔ̀ɔhâykɛ̀ɛt/ is a translation equivalent of the English verb *cause*, which has been established to display a negative semantic prosody (Stubbs 1995 *inter alia*). It will thus be interesting to determine whether or not /kɔ̀ɔhâykɛ̀ɛt/ also has a negative semantic prosody. Finally, my personal impression as a native speaker of Thai is that /chɔ̂ɔp/ is normally used in a negative context, and I would like to see whether it is the case that this word really has a tendency to occur in unfavourable environments, or that my native-speaker intuition on this point is misguided.

Each word was examined using two different methods. The first method was based on Sinclair's approach to semantic prosody as exemplified by his analyses of *naked eye*, *true feelings*, *brook*, and *place* (Sinclair 2004: 30-38). I examined 200 randomly-selected concordance lines for each word. Then, I identified the major patterns that exist around these words according to Sinclair's model of the extended unit of meaning, analysing colligation, collocation, semantic preference, and semantic prosody (since, in this approach, the semantic prosody of a unit of meaning cannot be considered independently of the other three phenomena.) The second method was based on Louw, Stubbs and Partington's approach. Here, I mainly looked at the words' statistically-strong collocates within a 4-left to 4-right span around the node, as suggested by Sinclair et al. (2004: 5); I classified the collocates as positive, negative or neutral as exemplified by the analyses in Louw (1993), Stubbs (1995), and Partington (1998; 2004). To measure collocational strength, I used the log-ratio measure of effect (Hardie forthcoming). Only items with log-ratio score higher than 3 that occur in at least five different texts were considered in my collocate analysis.

4 Results

The results of the collocate analysis for each word – based on the classification of individual collocates as positive or negative, without more detailed concordance analysis – are illustrated in Table 1.

Node	Collocate types/tokens		
	Positive	Negative	Neutral
/kreeŋcay/	2/35	5/91	9/142
/kɔ̀ɔhâykɛ̀ɛt/	8/316	44/1,744	20/1,587
/chɔ̂ɔp/	4/61	26/456	34/2,315

Table 1: The results of the collocate analysis

The concordance analysis, under Sinclair's approach shows that in many cases, /kreeŋcay/ is used on its own, that is, with only the colligations that are common to all verbs and without any easily classifiable pragmatic function beyond the core literal meaning of '(be) considerate (of)'. However, /kreeŋcay/ also appears in some fixed patterns of consistent combinations of collocation, colligations, and semantic preference, which we can describe as extended units of meaning in Sinclair's sense, as follows:

- /cà/, /yàk/, or /yàk cà/ + [verb group] + ([object/adverb]) + (/tɛ̀ɛ/) + /kɔ̂ kreeŋcay/ + ([person])
- This unit has a pragmatic function/ semantic prosody of 'refraining from performing an action due to consideration for someone'.
- [complete sentence-unit expressing imposition of hearer on speaker] + /mây tɔ̂ɔŋ kreeŋcay/ + /ná/, /ləəy/, or /rɔ̀ɔk/
- The unit has a pragmatic function/ semantic prosody of 'reduction of imposition' – specifically, the speaker asserts to the hearer that the previously-described imposition is not, in fact, an imposition on them.
- [action inconsiderate to another] + /yàaŋ/, /bɛ̀ɛp/, or /dooy/ + /mây kreeŋcay/ + ([person])
- The unit has a pragmatic function/ semantic prosody that expresses 'disapproval of behaviour'.

Similarly to /kreeŋcay/, /kɔ̀ɔhâykɛ̀ɛt/ and /chɔ̂ɔp/ are both used independently, with very general colligations and semantic preferences, but no clear extended pragmatic function – but also appear in some patterns that can legitimately be considered extended lexical units.

I argue that in the case of /kreeŋcay/, Sinclair's approach produces the better results, whereas for /kɔ̀ɔhâykɔ̀ɔt/, the positive vs. negative collocate analysis reveals more interesting results. For /kreeŋcay/, the positive vs. negative numbers are quite similar, and there are more neutral collocates, so the collocate analysis does not reveal much. The Sinclairian approach, on the other hand, produces some interesting patterns of /kreeŋcay/. The collocate analysis, however, reveals more intersting thing about the evaluation of causation, as the negative evaluation of causation is obvious from the big number of negative collocates. For /chɔ̂ɔp/, the analyses prove my intuition correct, as they show that /chɔ̂ɔp/ is normally used with a negative verb in a serial verb construction to indicate negatively-evaluated personal habits.

Acknowledgements

I would like to express my gratitude to Dr Andrew Hardie, my supervisor, for his invaluable comments on this paper.

References

Aroonmanakun, W. 2007. 'Creating the Thai National Corpus', *MANUSYA: Journal of Humanities* 13: 4-17.

Bednarek, M. 2008. 'Semantic preference and semantic prosody re-examined', *Corpus Linguistics and Linguistic Theory* 4 (2): 119-39.

Ebeling, S. O. 2014. 'Cross-linguistic semantic prosody: the case of 'commit', 'signs of' and 'utterly' and their Norwegian correspondences', *Oslo Studies in Language* 6 (1): 161-79.

Hardie, A (*forthcoming*) A dual sort-and-filler strategy for statistical analysis of collocation, keywords, and lockwords.

Hunston, S. 2007. 'Semantic prosody revisited', *International Journal of Corpus Linguistics* 12 (2): 249-68.

Louw, W. E. 1993. 'Irony in the text or insincerity in the writer? The diagnostic Potential of semantic prosodies', in M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*, pp. 157-76. Amsterdam: John Benjamins.

Partington, A. 1998. *Patterns and Meaning: Using Corpora for English Language Research and Teaching.* Amsterdam/Philadelphia: John Benjamin.

Partington, A. 2004. '"Utterly content in each other's company": semantic prosody and semantic preference', *International Journal of Corpus Linguistics* 9 (1): 131-56.

Partington, A. 2014. 'Evaluative prosody', in Aijmer, K and Ruhleman, C (eds.) *A Handbook of Corpus Pragmatics*, pp. 279-303. Cambridge University Press.

Sinclair, J. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Sinclair, J., Jones, S., Daley, R. and Krishnamurthy, R. 2004. *English Collocational Studies: The OSTI Report*. London: Continuum.

Stewart, D. 2010. *Semantic Prosody: A Critical Evaluation.* London: Routledge.

Stubbs, M. 1995. 'Collocations and semantic profiles: on the cause of the trouble with quantitative studies', *Functions of Language* 2 (1): 23-55.

Stubbs, M. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Whitsitt, S. 2005: 'A critique of the concept of semantic prosody', *International Journal of Corpus Linguistics* 10 (3): 283-305.

Zhang, W. 2009. 'Semantic prosody and ESL/EFL vocabulary pedagogy', *TESL Canada Journal,* 26 (2): 1-12.

Relative clause constructions as criterial features for the CEFR levels: Comparing oral/written learner corpora vs. textbook corpora

Yuka Takahashi
Tokyo University of Foreign Studies
takahashi.yuka.m0@tufs.ac.jp

Yukio Tono
Tokyo University of Foreign Studies
y.tono@tufs.ac.jp

1 Introduction

Since the Council of Europe officially announced the use of the CEFR for designing and evaluating foreign language syllabus and materials designs in each EU country in 2002, the use of the CEFR has been constantly expanding not only within Europe but also to the other parts of the world.

The CEFR is generic and language-independent. Thus it is underspecified as to what kind of grammar and lexis should be taught for each CEFR level. To supplement the framework, the procedure called Reference Level Descriptions (RLDs) has been undertaken, in which grammar points and lexical items are identified for each CEFR level.

Projects such as the English Profile Programme (EPP) (Hawkins and Filipovic 2012) use corpus data intensively in order to identify criterial features. The EPP especially is quite ambitious in the sense that they use both native and learner corpora to determine to what extent certain linguistic features serve as criterial for particular CEFR levels.

In the same vein, we have been investigating the nature of criterial features for Japanese learners of English, using our own corpus resources (Tono 2012; 2013). One of the features we focused on in this study is a relative clause (RC) construction, which is said to be one of the most difficult grammar items for learners of English (Hawkins and Buttery 2010) and also very frequently mentioned in SLA literature (cf. Ellis 2008: 562ff). By closely examining the state of acquisition of relative clauses, we hope to discover the path of identifying criterial features not just by quantitative, statistical methods, but also by looking at the process of acquisition in more detail.

This is a follow-up study of Takahashi & Tono (2014), which only looked at written learner corpora for the use of RCs. The present study examines both written and spoken learner corpora and the relationship between learner corpora and CEFR-based course book corpora.

2 Corpora used in the study

Table 1 shows the corpora used in this study. The Japanese EFL Learner (JEFLL) Corpus (Tono 2007) is comprised of written compositions by 10,038 Japanese secondary school (Year 7 to 12) students (669,304 running words). Originally the corpus was classified by school years, but for this RLDs project, the entire JEFLL Corpus has been re-classified into CEFR levels.

Learner corpora	Mode	Samples	Corpus size (sample n)
JEFLL Corpus	WR	Junior & senior high (all grades)	669,304 (10,038)
NICT JLE Corpus	SP	Adult	2,000,000 (1,281)

Table 1. Learner corpora used in the study

Both corpora were tagged for POS using TreeTagger. Extraction of relative clause constructions was done by writing pattern matching queries using regular expressions for the parts of speech of antecedents and each relative pronoun. The zero relative pronoun, which is common in producing contact clauses, was not covered in the present study.

In order to examine the instructional effects, two types of textbook data were analysed. One is a corpus of secondary school English textbooks published in Japan. The other is a corpus of ELT coursebooks based on the CEFR published in Europe. The former provides information about how often and in what order RCs are introduced in the school textbooks. The latter provides the data for the general order of presentation for different types of RCs.

All the instances of RCs were classified into the following categories:

- Categories based on the Noun Phrase Accessibility Hierarchy (NPAH) (Comrie & Keenan 1979) Hypothesis: S(subject)/DO(direct object)/IO(indirect object)/GEN(genitive)/OBL(oblique)/OCOMP(object of comparative 'than')
- Categories based on the SO Hierarchy Hypothesis (Hamilton 1994) : SS(=subject of the matrix clause & subject position of RC)/SO/OS/OO

Also each sentence was judged in terms of grammaticality and annotated for errors based on the following criteria:

- wrong selections of RCs
- resumptive pronouns
- wrong matrix positions

The present study aims to answer the following research questions:

- RQ1: Does the use of RCs increase along the CEFR levels, thus serving as criterial features?
- RQ2: Does the distribution of the use of RCs across the CEFR levels confirm the NPAH Hypothesis?
- RQ3: Does the distribution of the use of relative pronouns across the CEFR levels confirm the SO Hierarchy Hypothesis?
- RQ4: Does the mode of speech (spoken vs. written) affect the results of the above three questions?
- RQ5: Are there any similarities or differences in frequencies of different RCs between learner corpora and textbook corpora?

3 Results

The results show that basically the number of relative pronouns used in the JEFLL corpora was found to be increasing across the CEFR levels. Therefore, the first research question was confirmed.

Regarding the two hypotheses related to RQs 2 and 3, overall, while the SO Hierarchy Hypothesis was largely supported, the NPAH Hypothesis was partially supported due to the lack of evidence in GEN, OBL and OCOMP. These occurrences are also relatively infrequent in native corpora, compared to S and DO, so it seems that the results are reasonable.

The frequencies of RCs in the spoken learner corpora, NICT-JLE, were markedly lower than the written counterpart, which shows that the use of RCs in speech is more sensitive to the mode of speech (answer to RQ4). However, upper-intermediate speakers (B1 or B2 level) show more use of RCs in their speech, which suggests that the use of RCs can be a good criteria for distinguishing B level users from A levels.

Finally, the use of RCs in the textbook corpora (RQ5) is still underway and the results will be presented on the poster. We hope that the influence, either negative or positive, of textbooks used in Japan will be confirmed.

References

Comrie, B. and Keenan, E. 1979. Noun phrase accessibility revisited. *Language* 55: 649-664.

Ellis, R. 2008. *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Hamilton, R. 1994. Is implicational generalization unidirectional and maximal? Evidence from relativization instruction in a second language. *Language Learning* 44: 123-157.

Hawkins, J. and Buttery, P. 2010. Criterial features in learner corpora. *English Profile Journal* 1 (1), e5.

Hawkins, J. and Filipovic, L. 2012. *Criterial Features in L2 English*. Cambridge: Cambridge University Press.

Takahashi, Y. and Tono, Y. 2014. A learner corpus-based study on relative clause constructions as criterial features for the CEFR levels. A poster presented at TALC2014, Lancaster University.

Tono, Y. 2007. *Nihonjin 1-mannin no Eigo Corpus: JEFLL Corpus*. Tokyo: Shogakukan.

Tono, Y., Kawaguchi, Y. and Minegishi, M. (eds.) 2012. *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*. Amsterdam: John Benjamins.

Tono, Y. 2013. Automatic extraction of L2 criterial lexico-grammatical features across pseudo-longitudinal learner corpora: Using edit distance and variability-based neighbour clustering. In C. Bardel, C. Lindqvist & B. Laufer (eds), *L2 Vocabulary Acquisition: Knowledge and Use: New perspectives on assessment and corpus analysis* (pp.149-176). EuroSLA monographs. EuroSLA.

Aspectual discontinuity as a semantic-pragmatic trigger of evidentiality: Synchronic corpus evidence from Mandarin

Vittorio Tantucci
Lancaster University
v.tantucci@lancaster.ac.uk

1 Introduction

The present study is centered on the synchronic and diachronic interplay between aspectual **discontinuity** and **evidentiality** .

The former generally corresponds to the aspectual structure of experiential or existential perfects (i.e. Dahl & Hedin 2000; Portner 2003), or the so-called idle pasts (cf. Plungian & van der Auwera 2006:317) in which the current relevance at the utterance time of some previous event is not associated with some visible/verifiable results. While discontinuity (or anti-resultaivity) is not encoded through a specific grammatical construction in English, it is usually conveyed through 'anti-resultative' adverbials such as *once, before, in the past, previously* P or with impersonal usages of the present perfect as in *it has been* P (cf. Tantucci 2013: 219).

On the other hand, evidentiality is alternatively intended as the category referring to "the existence of a source of evidence for some information" (Aikhenvald 2004: 1) or the "encoding of the speaker's (type of) grounds for making a speech act" (Faller 2002: 2) or the communication of a piece of "acquired knowledge" (Tantucci 2013: 214). This work will focus on aspectual discontinuity as both a synchronic trigger of evidential strategies and a diachronic source of semasiological and grammatical reanalysis towards evidentiality. Namely, the semantic element of discontinuity will be shown to be pragmatically associated with some reliability behind the proposition, whereby the truthfulness of P is markedly 'at-issue' (cf. Faller 2002). More specifically, due to the inherent aspectual discontinuity of a construction, P is necessarily communicated either in the form of personal experience (as in the case of experiential perfects) or as a piece of interpersonally shared knowledge, also defined as **interpersonal evidentiality** (cf. Tantucci 2013, 2015).

The focus of the present survey is on the V-过 *guo* construction in Mandarin Chinese, which represents a perfect example of the synchronic and diachronic overlapping of the two domains of aspectual discontinuity and evidentiality. In fact, the V-过 *guo* construction is attested to have been through a process of grammaticalization from a previous experiential perfect stage (i.e. Cao 1995) towards a more recent interpersonal evidential (IE) employment (cf. Tantucci 2013, 2014, 2015).

Accordingly, based on the large qualitative annotation provided in Tantucci (2013), I will present new synchronic data which I gathered through the analysis of all the 862 occurrences of the chunk V-过 *guo* from the Lancaster Corpus of Mandarin Chinese (LCMC), a one-million word balanced corpus designed as a Chinese match of the Freiburg-LOB Corpus of British English (FLOB) (cf. McEnery & Xiao 2004). The findings of this survey will shed new light on the correspondence between specific written genres and the synchronic employment of V-过 *guo* either as a phasal (less-grammaticalized), an experiential or an interpersonal evidential (IE) marker. More broadly, this study will give an empirical account of the pragmatic motivations and textual environments contributing to the cognitive association between *lack of results at the moment of speech* (viz. aspectual discontinuity) and a piece of *shared knowledge within a community* (viz. interpersonal evidentiality).

2 Aspectual discontinuity and evidentiality: Diachronic evidence

Earliest usages of V-过 *guo* as an experiential perfect during the **唐** *táng* dynasty (618–907 A.C.) are attested to be limited to its co-occurence with animate subjects, mental verbs or verbs profiling the syntactic subject's personal experience in the past (cf. Cao 1995). However, Tantucci (2013:224–225, 2015) observes that during the **清** *Qīng* dynasty (636–1912 A.C.) V-过 *guo* will undergo a further stage of semantic and grammatical reanalysis, as it will start to co-occur with dummy subjects, in subjectless or impersonal constructions with a new interpersonal evidential (IE) meaning. More specifically, functioning as an IE, V-过 *guo* will be no longer employed as an aspectual marker of past experience, but rather used to problematize the reliability of P as a piece of knowledge shared by the SP/W together with a general 3rd party in society, paraphrasable as: *it is known that* P.

I argue that the semasiological shift from 'past experience' to 'shared knowledge' is precisely triggered by the inherent discontinuous aspectual structure of V-过 *guo*, which depending on the context, the textual environment and the degree of grammaticalization of 过 *guo* as a particle, functions as a bridging element from a mere aspectual to a new evidential reading. Compare the two examples below:

看 文字 须 仔细，虽是 旧 曾 看 过，重温 亦
kàn wénzì xū zǐxì suíshì jiù céng kàn guo chóngwēn yì
see character must careful although old once see EXP review also
须 仔细。
xū zǐxì
must careful
'When you look at a character you must be attentive, even if it is one that you saw before, you still have to be attentive.'
朱子语类 *zhūzǐyǔlèi*，(Cao 1995: 41)

在 他们 乡 里， 发生 过 这样 一 件 事: […]
zài tāmen xiāng lǐ, fāshēng guo zhèyàng yī jiàn shì
in they village inside, happen IE such one CLASS thing
'(It is known that) **something happened** in their village: […]'
PKU-CCL Gǔ Jīn Qíng Hǎi Narrative 1915

Interestingly Plungian & van der Auwera (2006: 317) discuss in detail the typological relevance of discontinuity/ anti-resultativity in the aspectual and tense systems of world languages, whereby "discontinuous past appears to be represented in many genetically unrelated languages of different areas". Consider the example below of the discontinuous marker *-na* from Futunan:

na koi su'a le li'ua
DP CONT flow SP river
'At that time, the river still used to flow'.
(Moyse-Faurie 1993: 210)

In the case above, SP/W is stating P based on a piece of markedly interpersonal knowledge (cf. Moens & Steedman; Rubovitz 1999; Portner 2003 on the relationship between evidential reasoning and aspectual discontinuity). Interestingly enough, the example above presents striking similarities with the IE employment of 过 *guo* in Mandarin, in Sinitic languages in general and other evidential systems such as Persian where discontinuous and remote past forms also convey evidentiality (cf. Lazard 1999: 99). The following example is from Taiwanese Southern Min (2001:65):

遮 識 出 過 水泉
chia bat chut koè chúi-chôa
here EVD appear EVD spring
'There used to be a spring here.'
(Chappell 2001: 65)

As can be noted form the occurrence above, the IE meaning conveyed by '**識** *bat* -V- 过 *guo*' regards an event or situation the evidence of which is presented as a piece of knowledge shared by SP/W and 3rdP, again paraphrasable as *apparently, as is known, it seems* and similar IE adverbials (cf. Tantucci 2013: 223).

3 From experience to evidence: A synchronic corpus study of the pragmatics of V-过 *guo* and specific text types

What will emerge from the synchronic survey from the LCMC is that experiential usages of V-过 *guo* are more prototypically employed in narrations in comparison with evidential usages of the chunk. On the other hand, the evidential employment of V-过 *guo* in Press registers will appear significantly higher than the ones of experientials ($p < 0.008$). Additionally, evidential functions of the chunk will be shown to be significantly higher than experientials in the factual/ academic registers of the LCMC ($p < 0.039$).

In the final analysis, the diachronic shift from personal experience to interpersonal knowledge will be shown to be reflected synchronically in the different genres of the LCMC. These data will further support the claim of a semantic-pragmatic continuum from aspectual discontinuity to interpersonal evidentiality depending on the register and the textual environment in which the construction tends to occur.

References

Aikhenvald, Alexandra Y. 2004. *Evidentiality*. Oxford: Oxford University Press.

Cao, Guang Shun. 1995. *Jindai hanyu zhuci [The auxiliary particles of Modern Chinese]*. Beijing: Yuwen chubanshe [Language & Culture Press].

Chappell, Hilary. 2001. "A Typology of evidential markers in Sinitic languages." In *Chinese grammar: Synchronic and diachronic perspectives*, edited by H. Chappell, 56–85. New York: Oxford University Press.

Dahl, Östen, and Eva Hedin. 2000. "Current relevance and event reference." In *Tense and aspect in the languages of Europe*, edited by Ö. Dahl, 385–402. Berlin: Mouton de Gruyter.

Faller, Martina. 2002. "Semantics and pragmatics of evidentials in Cuzco Quechua."PhD, Stanford University, Standford University.

Lazard, Gilbert. 1999. "Mirativity, evidentiality, mediativity, or other?" *Linguistic Typology* 3 (1):91–109.

McEnery, Anthony, and Zhonghua Xiao. 2004. "The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study." *Religion* 17:3–4.

Moens, Marc, and Mark Steedman. 1988. "Temporal ontology and temporal reference." *Computational linguistics* 14 (2):15-28.

Moyse-Faurie, Claire. 1993. *Dictionnaire futunien-francais avec index francaise-futunien: Avec index français-futunien*. Leuven: Peeters Publishers.

Plungian, Vladimir A., and Johan van der Auwera. 2006.

"Towards a typology of discontinuous past marking." *Language Typology and Universals* 59 (4):317–349.

Portner, Paul. 2003. "The (temporal) semanstics and (modal) pragmatics of the perfect." *Linguistics and Philosophy* 26 (4):459–510.

Rubovitz, Tali. 1999. "Evidential-existentials: The interaction between discourse and sentence structure." *Journal of pragmatics* 31 (8):1025–1040.

Tantucci, Vittorio. 2013. "Interpersonal Evidentiality: The Mandarin V-过 guo construction and other evidential systems beyond the 'source of information'." *Journal of Pragmatics* 57:210–230.

Tantucci, Vittorio. 2014. "Epistemic inclination and factualization: A synchronic and diachronic study on the semantic gradience of factuality." *Language and cognition* FirstView. doi: 10.1017/langcog.2014.34.

Tantucci, Vittorio. 2015. "Traversativity and grammaticalization: The aktionsart of 过 guo as a lexical source of evidentiality." *Chinese Language and Discourse*:57–101.

'Why are women so bitchy?': Investigating gender and mock politeness

Charlotte Taylor
University of Sussex / Lancaster University
charlotte.taylor@sussex.ac.uk

1 Introduction

The question posed in the title comes from a conversation within my dataset and it is one that I intend to address in this paper, although probably not in the way that author intended. What I am interested in is the use of the label *bitchy*, and other terms, to describe women's behaviours and how these relate to labels used to describe men's behaviours. More specifically, I focus on behaviours which involve some kind of mismatch in politeness, as in:

> I usually say something if someone doesn't thank me to be honest - a sarcastic "no problem" might remind them to be polite next time. (*example from the mumsnet corpus*)

The aims of this paper are twofold: first, to investigate the extent to which perceptions of gender and mock polite behaviour correlate with actual usage and, second, to explore the limits of corpus linguistics in this kind of pragmatic enquiry.

2 Mock politeness

Mock politeness is used here as a broad term intended to encompass all verbal behaviours in which when there is a politeness mismatch leading to an implicature of impoliteness. It is a subset of Culpeper's (2011) category of implication impoliteness and draws on Leech's (1983/2014) Irony Principle.

In previous research, I have found that the following labels are used to describe mock politeness in conversation: *bitchy*, *biting*, *caustic*, *condescending*, *cutting*, MAKE FUN, MOCK, *passive aggressive*, *patronising*, PUT DOWN, *overly polite*, TEASE. These labels will be used here to identify mock polite behaviours, adopting a first-order approach.

3 Gender and mock politeness

Previous research into the relationship between gender and mock politeness has focussed on behaviours labelled as *sarcastic* and *patronising/condescending*

More specifically, studies of gender and sarcasm have tended to focus on whether men or women are more likely to use sarcasm and whether reception of

sarcasm differs according to the addressee or target of the sarcastic behaviour. To date, there is a consensus amongst researchers that men are more likely to use sarcasm than women. However, these findings are frequently based on self-reports in which participants are asked whether they are sarcastic/use sarcasm (e.g. Rockwell & Theriot 2001; Ivanko et al. 2004; Dress et al. 2008; Colston & Lee 2004). From a corpus pragmatic perspective, this is deeply worrying because what is actually being assessed is perception of use, not actual use. And, equally worrying from a corpus semantic/pragmatic perspective, what is being assessed is identification with a particular label, and not performance of a particular behaviour.

Thus, what I wanted to verify in this study was a) whether the range of terms used for describing male and female performance varied according to gender and b) whether the actual behaviours varied. In other words, is a woman being *bitchy*, doing the same thing as a man being *sarcastic*?

4 Corpora and tools

The main corpus used in this study comprises approximately 61 million tokens of forum interactions from the UK website mumsnet.com. This source was selected because it allows access to 'everyday' or 'conversational' comments on im/politeness, and thus offers a 'wide range of different kinds of mundane everyday interactions beyond what a single researcher might realistically collect' (Haugh 2014: 83).

The corpus was downloaded using BootCaT using a wide range of im/politeness related search terms and was interrogated using Sketch Engine (Kilgarriff et al 2004), WordSmith Tools (Scott 2008) and the Collocational Network Explorer (Gullick and Lancaster University 2010).

5 Process: A three-pronged approach

I take a three-stage approach to the investigation, combining the methodologies Corpus-Assisted Discourse Studies (Partington 2004; Partington et al. 2013) with an experimental approach more common in psychological studies of irony:

In the first stage, I exploit the ability of corpus tools to handle very large amounts of data by using collocation analyses to investigate the mock politeness labels for gender associations. This stage showed how several meta-pragmatic labels are more strongly associated with men/women and boys/girls.

In the second stage, I make use of more typical corpus pragmatic tools by using a heavily marked up and annotated sub-corpus. This sub-corpus is made up of the behaviours which were referred to using the mock politeness labels and the added information includes information such as the gender of the speaker and type of im/politeness mismatch which occurs. This is then used to verify to what extent the behaviours which are associated with particular genders are similar or different. The findings from this stage show that there were many similarities between the differently labelled behaviours.

In the third stage, I step away from the corpus analysis because this is the point I feel marks the limitation of the direct corpus analysis. In this stage, I take the examples of mock politeness analysed in the previous stage and manipulate the gender of the speakers in these examples so they can be used in an experimental extension. The original and modified examples will be given to participants who will be asked to describe the behaviour using one of a set of meta-pragmatic labels. The aim of this stage is to verify to what extent the choice of a particular label is influenced by the perceived gender of the speaker.

6 Conclusions

Two main aspects are addressed in this paper. In the first I analyse the relationship between gender and mock politeness at the level of perception and evaluation, following in the footsteps of Baker's (2014) work on gender using corpora. In the second, more methodologically-focused aspect, I examine the ways in which corpus linguistics may be combined with other approaches in order to complete and complement analyses.

References

Baker, P. 2014. *Using Corpora to Analyze Gender*. London & New York: Bloomsbury.

Colston, H. L., & Lee, S. Y. 2004. Gender differences in verbal irony use. *Metaphor and Symbol*, 19(4), 289-306.

Culpeper, J. 2011. *Impoliteness: Using Language to Cause Offence*. Cambridge University Press.

Dress, M.L., R.J. Kreuz, K.E. Link and G.M. Caucci. 2008. Regional Variation in the Use of Sarcasm. *Journal of Language and Social Psychology* 27: 71.

Gibbs, R. W. 2000. Irony in talk among friends. *Metaphor & Symbol* 15 (1&2): 5-27.

Gullick, D. & Lancaster University 2010. Collocational Network Explorer (CONE) Available from https://code.google.com/p/collocation-network-explorer/

Haugh, M. 2014. Jocular mockery as interactional practice in everyday Anglo-Australia conversation. *Australian Journal of Linguistics* 34, 1: 76-99.

Ivanko, S. L., Pexman, P. M., & Olineck, K. M. 2004. How sarcastic are you? Individual differences and verbal irony. *Journal of Language and Social*

Psychology, 23(3), 244-271.

Kilgarriff, A., P. Rychly, P. Smrz and D. Tugwell. 2004. The Sketch Engine. Proceedings of EURALEX 2004, Lorient, France; pp 105-116 http://www.sketchengine.co.uk

Leech, G. 1983. *The Principles of Pragmatics*. London & New York: Longman.

Leech, G. 2014. *The Pragmatics of Politeness*. Oxford: OUP.

Partington, A. 2004. Corpora and discourse, a most congruous beast. In A. Partington, et al. (eds.) Corpora and Disocurse. Bern: Peter Lang, pp. 11–20.

Partington, A., Duguid, A. and Taylor, C. 2013. Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS). Amsterdam: John Benjamins.

Rockwell, P. & E.M. Theriot 2001. Culture, gender, and gender mix in encoders of sarcasm: A self-assessment analysis. *Communication Research Reports* 18(1):44-52.

Scott, M., 2008, *WordSmith Tools version 5*, Liverpool: Lexical Analysis Software.

Facebook in the Australian News: a corpus linguistics approach

Penelope Thomas
University of Sydney

ptho2197@uni.syd.edu.au

With the launch of its newsfeed application 'Paper' in February, 2014, Facebook became a news provider. It is now a social networking site that also selects and curates news via an in-house editorial team. All Facebook functionality is built into the app so that social networking, news reading and news sharing can be seamless activities. Research on Facebook to date considers its applications, its prominence as a leading social networking site (Boyd and Ellison 2008), and the language and behaviours of its users (Ellison et al. 2007; DeAndrea et al. 2010; Bouvier 2012). The focus of this project, however, is the way a social media company like Facebook has fast become a mainstream news provider. The relationship between this emerging platform and the existing one of traditional news is the interest that spurred this research.

To explore this topic, the current project aims to provide a linguistic description of news discourse around Facebook itself and a focused case-study of news values using a corpus linguistics approach. It makes use of a 101,900 word 'specialised' corpus (Hunston 2002:14) called Facebook News Corpus, or FNC, consisting of Australian news text that appeared around three main events in the company's history: 1) the launch of Facebook in Australia; 2) the listing of Facebook Inc. on Nasdaq; and 3) the introduction of Graph Search. It applies the 'discursive framework' of Bednarek and Caple (2012a/b; 2014) and is the first study to systematically test discursive news values analysis on a specific topic.

News values have been defined from a variety of perspectives as: 'factors influencing the flow of news' (Galtung and Ruge 1965:64); 'an ideological code' (Hartley 1982:81); 'the cognitive basis for decisions about selection, attention, understanding, representation, recall, and the use of news information in general' (van Dijk 1988b:119); 'intersubjective mental categories' (Fowler 1991:17); '[an] often unconscious criteria by which newsworkers make their professional judgements' (Bell 1991:155-156); and, most recently, as having a 'discursive' dimension in being 'established by language and image in use' (Bednarek and Caple 2012:44-45). Using the latter definition, key research questions about news values for this project are: How is 'newsworthiness' (news values)

constructed in news stories around Facebook compared to general news? How useful are corpus linguistics techniques in applying discursive news values analysis and answering this question?

The approach here is primarily inductive, using ranked word frequency lists as a starting point. The two following phases of analysis draw on both quantitative and qualitative methods. Phase 1 systematically tests Bednarek and Caple's discursive framework (2012a/b;2014), and compares the Facebook News Corpus (FNC) with a general news corpus, BNC-Baby 2.2 Newspapers, in terms of keyword analysis checked against distribution (Gries 2008), and 'diachronic' change (Baker 2006:29). Phase 2 looks at collocates of 'Facebook' and 'Zuckerberg'.

Phase 1 of analysis has gathered some potentially significant findings. In comparing FNC with BNC-Baby, the keyword tool provided insight into the 'aboutness' (Scott and Tribble 2006:53) of news discourse around Facebook by showing which words are more frequent in FNC when compared with a general news corpus. Not surprisingly, words about social media are prevalent in the FNC top 100 keyword list. These can be clearly categorised into different semantic groups: the entities of social media and the internet; people and places, including those associated with social media companies and words used to describe them, physical geographical locations and social media users; activities around social media and internet use; and words about the commercial aspect of social media.

Although keyword analysis is useful in providing a general overview of corpus data, it does not necessarily indicate the construction of news values. For example, the positive keyword ranking of *privacy* (ranked 26), *information* (ranked 41) and *friends* (ranked 22) signifies these are topics that strongly relate to news about Facebook, yet closer analysis is needed to gain a more accurate picture of how these items are actually being used. Using the discursive framework (Bednarek and Caple 2012 a/b; 2014), the top 100 keywords are reclassified according to their potential as pointers to news values: Negativity, Timeliness, Proximity, Superlativeness, Eliteness, Impact, Novelty, Personalisation, and Consonance.

At this stage of the framework, assigning items to news values is still hypothetical, so contextual analysis of concordance lines allows for a more insightful evaluation. In the case of *privacy*, concordancing shows that it is predominantly used around language that clearly fits within the Negativity news value (Galtung and Ruge 1965:69). Words in proximity can also be semantically grouped into: 1) words that describe the impact on the individual Facebook user: destroy, fears, violating, serious invading, major invading, eroded, threat, assault, forced; 2) words that refer to the needs and potential reaction of users to issues: protection, containment, warned, cautious, risked, revolt, demonstrate, backlash, tighten; and 3) words that refer to legal terms and government-related issues: officer, policies, policy, right, breeches.

These and other findings are part of an effort to construct a linguistic description of news text about Facebook. This information will be useful for future research on the role of social networking sites and changing media practices (Dwyer 2010), given some researchers consider Facebook as a primary news source for its users.

References

Baker, P. 2006. *Using Corpora in Discourse Analysis*. London/New York: Continuum.

Bednarek, M. and Caple, H. 2012a. *News Discourse*. London/New York: Continuum.

Bednarek, M and Caple, H. 2012b. ""Value added": Language, image and news values". *Discourse, Context, Media* 1: 103-113.

Bednarek, M. and Caple, H. 2014. "Why do news values matter? Towards a new methodological framework for analyzing news discourse in Critical Discourse Analysis and beyond'. *Discourse & Society* 25/2: 135-158.

Bell, A. 1991. *The Language of News Media*. Oxford: Blackwell.

Bouvier, G. 2012. "How Facebook users select identity categories for self-presentation". *Journal of Multicultural Discourses* 7 (1): 37-57.

Boyd, D. and Ellison, N. 2008. "Social Network Sites: Definition, History, and Scholarship". *Journal of Computer-Mediated Communication* 13 (1): 210-230.

DeAndrea, D. C., Shaw, A. S., and Levine, T. R. 2010. "Online language: The role of culture in self-expression and self-construal on Facebook". *Journal of Language and Social Psychology* 29 (4): 425-442.

Dwyer, T. 2010. *Media Convergence*. Maidenhead: University Open Press.

Ellison, N., Steinfield, C. and Lampe, C. 2007. "The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Networking Sites". *Journal of Computer-Mediated Communication* 12 (4), 1143-1168.

Fowler, R. 1991. *Language in the News: Discourse and Ideology in the Press*. London/New York: Routledge.

Galtung, J. & Ruge, M. 1965. "The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers". *Journal of Peace Research* 2 (1): 64- 90.

Gries, St. Th. 2008. "Dispersions and adjusted

frequencies in corpora". *International Journal of Corpus* Linguistics 13 (4): 403-37.

Hartley, J. 1982. *Understanding News*. London: Methuen.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

van Dijk, T. 1988. *News as Discourse*. Hillsdale: Lawrence Erlbaum.

Linguistic feature extraction and evaluation using machine learning to identify "criterial" grammar constructions for the CEFR levels

Yukio Tono
Tokyo University of Foreign Studies
y.tono@tufs.ac.jp

1 Background

The purpose of this study is twofold. One is to find an effective method of identifying grammatical and lexical features which can serve as criteria for distinguishing a given level of the Common European Framework (CEFR) from other levels. The second aim is to prepare a list of grammatical and lexical items for each CEFR level as an inventory.
This is part of the on-going research for the CEFR-J project (Negishi, Takada & Tono 2012). The CEFR-J (Japan) is an adapted version of the CEFR for English language teaching in Japan. The CEFR-J closely follows the original CEFR in its basic design, consisting of the six common reference levels (A1 to C2), with further branching in lower levels (e.g. three sub-levels under A1 (A1.1 to A1.3) and two for each level from A2 to B2, respectively). The CEFR-J went through a series of similar validation processes following the original CEFR, such as sorting exercises of can do descriptors by a group of EFL teachers and the can-do questionnaire survey by more than 5,000 students with various levels of proficiency, whose results were used for empirically validating the order of difficulties among the "can do" descriptors using the Rasch model.

The Ministry of Education, Culture, Sports, Science and Technology, Japan (MEXT) is keen to apply the CEFR-like concept of "can do" descriptors as performance objectives in EFL teaching in Japan for the next revision of the national curriculum called the Course of Study.

The goal of the CEFR-J project is to support the government for their development of the Course of Study and textbook companies for their production of ELT course books based on the Course of Study by preparing an inventory of grammar and vocabulary points describing reference levels specified in the CEFR-J. Such resources are available already for the original CEFR, e.g. the *Core Inventory for General English* (North et al. 2010), but they are not attested against real language use data. The English Profile () aims to provide reference level descriptions similar to ours using the Cambridge Learner Corpus (Hawkins and Filipovic 2012), but our project focuses on Japanese learners

of English with special reference to the CEFR-J further-branching levels. To this end, the CEFR-J project aims to obtain such an inventory by a corpus-based approach. The project is currently called the CEFR-J RLD project.

2 Corpora compiled for the project

In order to identify language features describing English CEFR levels, two types of corpora have been compiled: textbook corpora and learner corpora. Textbook corpora consist of CEFR-based ELT course books. The textbooks were carefully selected in such a way that only those published in Europe after the formal launch of the CEFR in 2001 were chosen and that their syllabuses are based on CEFR levels and descriptors (see Table 1). Altogether 95 books from 4 publishers covering A1 to C2 were selected. Due to the copyright issues, this resource is available for internal use only for the moment.

Name	CEFR-level	Skills	Corpus size
Textbook Corpus	A1 to C2	All skills	2,800,000 (95 titles)

Table 1: Textbook corpora

Since this inventory is designed for Japanese learners of English, the project also collected various learner corpora (see Table 2).

Learner corpora	Mode	Samples	Corpus size (sample n)
JEFLL Corpus	WR	Junior & senior high (all grades)	670,000 (10,036)
NICT JLE Corpus	SP	Adult	2,000,000 (1,281)
MEXT Data	WR/SP	Junior High 3rd	100,000 (2,000)
GTECfS Writing Corpus	WR	Senior High 1-3	2,500,000 (30,000)

Table 2: Learner corpora used for the study

It is true that the Cambridge Learner Corpus is much bigger than ours, but the CEFR-J RLD project only focused on Japanese learners of English, and in terms of a single homogeneous learner group, our data set is more finely tuned and one of the largest among available learner corpora and covers different proficiency levels (novice to intermediate) and modes of speech (spoken and written).

3 Feature extraction

The next step is to decide what grammatical and lexical features to extract from CEFR-classified data. The feature list was deliberately made as exhaustive as possible, because the main purpose of the project is to prepare the inventory of all the learning items to be introduced or taught at each CEFR or CEFR-J level. For assessment purpose, on the other hand, the list could be much shorter and limited only to some features which are significantly more useful to discriminate the levels.

To this end, a list of all the grammar items introduced at the first six years of instructions (junior high and senior high schools) was prepared and query patterns were written using POS tags and regular expressions. Altogether 144 grammar items were combined with 14 sentence patterns (e.g. declarative, interrogative, negative, wh-questions) and altogether 1,320 patterns were prepared (Tono and Ishii 2014).

In order to extract grammar patterns from different skill sections, each activity in the textbooks was carefully annotated for skill types specified in the CEFR (listening, spoken production, spoken interaction, reading, and writing). The underlying assumption is that the characteristics of vocabulary and grammar used in receptive vs. productive tasks should be different, and that profiling such information as different types of input is important. For the output learner data, both spoken and written corpora were prepared, but this present study only examined written sections of learner corpora.

4 Evaluation of feature extraction

Since some of the grammar patterns were quite complicated, the accuracy of query patterns was evaluated by randomly sampling some patterns and compare the results against human manual counts. Table 3 shows some examples.

Grammar items	Precision	Recall	F
Subject-position relative clause	0.98	0.74	0.85
Present perfect (@have been)	1.00	1.00	1.00
Get + past participle	0.99	0.55	0.70
It … to V	1.00	0.98	0.99
Subjunctive past	1.00	0.14	0.24

Table 3: Accuracy of query patterns

Some patterns were low in accuracy for various reasons. The pattern "get + past participle", for instance, could not retrieve "get used to" due to irregular POS information. Also subjunctive past was low because such phrases as "I wonder if …" or "even if …" were not included. The query syntax will be improved in the future based on the

evaluation of these preliminary search results.

5 Machine learning

The current phase of the project has tested several different machine learning algorithms in order to evaluate the quality of classifiers and additional resources produced. In this pilot research, 144 grammar items were extracted for declarative and negative sentences, which produced the frequency matrix of altogether 228 items across CEFR-classified textbook corpora.

Three well-known machine learning algorithms were tested: J48, Support Vector Machine (SVM), and Random Forest (RF). Weka 3.6.11 (Hall et al. 2009) was used for the analysis. Table 4 shows the results of each algorithm and its F-measure in the classification task of CEFR texts.

		Precision	Recall	F-measure
J48	A1	0.929	1	0.963
	A2	0.591	0.619	0.605
	B1	0.379	0.407	0.393
	B2	0.35	0.292	0.318
	C1	0.1	0.111	0.105
SVM	A1	0.538	0.538	0.538
	A2	0.48	0.571	0.522
	B1	0.458	0.407	0.431
	B2	0.435	0.417	0.426
	C1	0.3	0.333	0.316
RF	A1	0.8	0.923	0.857
	A2	0.65	0.619	0.634
	B1	0.548	0.63	0.586
	B2	0.556	0.625	0.588
	C1	0	0	0

Table 4. The performance of classifiers

J48 (the implementation of C4.5) showed excellent results especially for lower levels (A1 and A2), while Support Vector Machine outperformed J48 for intermediate levels (B1 through C1). Random Forest showed high F-measures, compared with the other two, for both lower and intermediate levels. However, it completely failed to classify C1-level texts. RF performed best in the present author's previous study with learner data (Tono xx), thus it will be promising to use RF for the overall analysis if the causes of the low F-measure for C1 were identified.

6 Learning from attribute weights

One of the advantages of using classifiers is that we can obtain the valuable information about relative weights of attributes or predictive variables used for the classification. This information will help decide which attributes are most useful for classifying texts.

Table 5 shows the attribute weights produced by SVM.

```
Classifier for classes: A1, A2

        -0.1237 * (normalized) 1-1
  +     -0.4007 * (normalized) 1-2
  +     -0.6031 * (normalized) 1-3
  +      0.1508 * (normalized) 1-4
  +     -0.2654 * (normalized) 1-5
  +      0.1305 * (normalized) 1-6
  +      0.2963 * (normalized) 1-10
  +     -0.3743 * (normalized) 1-11
  +      0.043  * (normalized) 1-12
  +      0.1917 * (normalized) 1-13
  +     -0.5896 * (normalized) 1-14
  +     -0.4946 * (normalized) 1-15
  +     -0.4977 * (normalized) 1-16
  +     -0.1117 * (normalized) 1-17
  +     -0.1549 * (normalized) 1-18
  +     -0.4934 * (normalized) 1-19
  +     -0.0685 * (normalized) 1-20
```

Table 5. Attribute weights by SVM

In Table 5, each grammar item (e.g. 1-1 through 1-20 among 122 items) has either positive or negative weights. The positive value means the given attribute contributed to the classification into A2 level. The negative, vice versa. Table 6 shows some of the specific grammar items which were found useful for classifying CEFR level texts from one another.

Classify	Useful attributes
A1 vs. A2	A1: He [She] is NP; It is NP; It is not NP A2: when-clause; SVOC (negative)
A1 vs B1	A1: It is PP B1: be able to; how NP+VP; Present perfect continuous
A1 vs B2	A1: He [She] is NP; He [She] is ADJP B2: how NP+VP; V + it + ADJ to do …

Table 6. Some useful grammar items for CEFR-level classification based on SVM attribute weights

It is the goal of the CEFR-J RLD project to produce all the weights of grammar items as attributes for classifying CEFR texts both for textbook and learner corpora, and prepare the inventory of grammar with the weight information. This will be of great help in designing the sequence and grading of these grammar items along the CEFR levels. The same kind of evaluation will be made for lexical profile measures (fluency, complexity and accuracy) besides grammar items. The entire inventory will be released by March 2016.

References

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, Volume 11, Issue 1.

Hawkins, J. and Filipovic, L. 2012. *Criterial features in L2 English: Specifying the Reference Levels for the Common European Framework*. Cambridge: Cambridge University Press.

Negishi, M., Takada, T. and Tono, Y. 2012. A progress report on the development of the CEFR-J. *Studies in Language Testing* 36: 137-165. Cambridge: Cambridge University Press.

North, B., Ortega, A., and Sheehan, S. 2010. Core Inventory for General English. British Council/ EAQUALS. Available at http://englishagenda. britishcouncil.org/sites/ec/files/books-british-council-eaquals-core-inventory.pdf

Tono, Y. 2013. Criterial feature extraction using parallel corpora and machine learning. In Diaz-Negrillo, A., Ballier, N., and Thompson, P. (eds.) *Automatic treatment and analysis of learner corpus data.* Amsterdam: John Benjamins.

Tono, Y. and Ishii, Y. 2014. Extraction of criterial features for the CEFR: Automatic grammar extraction and evaluation. Presentation given at JAECS 40, Kumamoto Gakuen University, 4 October, 2014.

A corpus analysis of EU legal language

Aleksandar Trklja
University of Exeter

1 Introduction

This paper reports results of a study conducted as part of the project 'Law and Language at the Court of Justice of the European Union' funded by the European Research Council. The Court of Justice of the European Union (CJEU) is responsible for the creation of EU case law. It interprets EU law and makes sure that it is applied in the same way in all EU countries (e.g. Paunio, 2011). Judgments produced by the CJEU are drafted in French and then translated in other official EU languages (24 at the moment). The purpose of the study is to identify key linguistic features of EU case law and to investigate the impact they might have on the content of EU case law.

2 Corpora and tools

Two types of corpora were compiled for the purposes of the present study. Our first corpus (CJEU corpus) was created by compiling 1140 judgments produced in English, French, German and Italian (more languages will be included in future). These judgments together with other types of legal documents form the core of EU law. We also compiled a reference corpus (REF corpus) that contains all online available case law judgments produced by seven EU Constitutional or Supreme Courts in the same four languages. Both corpora are tagged using TreeTagger (Schmid 1994) and indexed in CWB (Evert, 2005). In addition, various reference corpora available in the Sketch Engine (Kilgarriff et al., 2014) were used.
In addition to Sketch Engine and CWB tools we also used Collocate and WordSmith Tools (Scott, 2008). Besides, we also wrote several shell and Python scripts for specific tasks (e.g. to compare collocation framework of lexical items, to extract specific text sections from judgments).

3 Research aims

The general purpose of the study is to identify and describe key linguistic features common to different language versions of CJEU judgments. In particular, we wanted to provide evidences for the following assumptions proposed in previous research (McAuliffe, 2009):

- the language used in CJEU judgments differs from other legal languages,

- the translation process influences the content of judgments in different languages,
- the language of CJEU judgments is highly repetitive.

4 Findings

A comparative analysis of collocation profiles indicates that there are differences in the use of same multi-word lexical items in judgments from the CJEU and REF corpus. For example, in the English version of CJEU judgments <competent to> collocates with 21 verbs, whereas in judgments from the REF corpus (which is about three times bigger) the same expression occurs with 13 such collocates. Only five collocates occur in both corpora in this context. More than half (62%) of verbal collocates of <competent to> found in CJEU judgments do not occur in REF or in BNC (e.g. <adjudicate>, <apply>, <conclude>, <entertain>, <establish>, <impose>, <provide>, <raise>, <re-examine>, <reopen> and <rule out>). This figure is lower for judgments from REF (30%). The collocation profile of this and other items indicate that the EU case law contain expressions that are not typical of other text types in English.

To investigate similarities and differences between multi-word expressions we also studied functions of 5-grams in the two corpora. For example, in both corpora we find 5-grams that refer to different types of courts (e.g. <in the court of appeal>, <the court of first instance>) or those that have a textual function (e.g. <in so far as it>, <in the light of the>, <on the basis of the>). However, the latter type of 5-grams are more typical of the CJEU judgments. On the other hand, only in judgments from REF 5-grams express personal stance (e.g. <i do not think that>, <it seems to me that>). In CJEU judgments more typical are impersonal expressions (e.g. <it is clear from the>, < it is apparent from the>). The most frequent 5-gram in REF expresses rhetorical politeness (e.g. <my noble and learned friend>) and this type of language is completely absent in CJEU judgments. Finally, the linguistic units expressing obligation are more typical of CJEU judgments (e.g. <it is necessary to examine>, <it should be noted that>).

For some of multi-word expressions it can be said that they result from the translation process and/or languages-in-contact situation. Here are some examples of expressions created through translation from French into English: <concentration compatible with> from <concentration compatible avec>; <implementation of the concentration> from <réalisation de la concentration>; <competitive constraint> from <contraintes concurrentielle>; or <competitive neutrality> from <neutralité concurrentielle>.

Formulaic expressions can be found in texts from both CJEU and REF but CJEU judgments have a higher degree of formulaicity. For example, on the average 46% of the text of CJEU judgments in English consists of repetitive expressions which are at least five words long. The figure for UK judgments is 37% and for Irish judgments 39%. In the German version of these CJEU judgments repeated expressions make up on the average 37% of the text and in judgments produced by the German and Austrian Constitutional Courts 33% and 23% respectively. The average length of repeated expressions also tends to be higher in CJEU judgments (in English, French and German versions of CJEU judgments they are 60 words long and in the REF corpus 30 words long).

In addition to this direct type of repetition there are also semantic repetitions realized through synonyms-like expressions. These phrases are identified in a parallel corpus of CJEU judgments. We assume that items from language A that corresponds to items from language B and are used in the same context have the same function. Thus, in our context we find that <it must be borne in mind>, <it must be recalled>, <it must be pointed out>, <it should be noted> or <it should be observed> are interchangeable and therefore have the same function because they meet the above criteria. The difference which is usually made in the literature (e.g. Coates, 1983) between a stronger (<must>) and weaker <should> notion of necessity is here ignored. Similarly, following what can be found in the dictionaries consulted and in the results yielded by the Sketch Engine tools Thesaurus and Sketch-Diff the verbs <bear in mind>, <recall>, <point out>, <note> and <observe> do not tend to be used as synonyms generally in English.

5 Conclusion

In our study we identified some of key linguistic features of case law judgments produced by the CJEU. Our findings demonstrate that these judgments regardless of language versions tend to be formulaic and repetitive, that they contain expressions typical of EU case law and that some of these expressions are created through translation. Our results support findings from interview data published in earlier studies (McAuliffe, 2009; McAuliffe, 2010).

Given the fact that EU case law is typically drafted by non-native speakers it seems that formulaic language serve as a useful tool here. Drafters of CJEU judgments seem to think in terms of pre-fabricated expressions. Repetitive expressions become embedded in the developing legal order and it follows that the language shapes the concepts

being developed and also helps to embed EU law as a system.

References

Coates, J. 1983. The semantics of the modal auxiliaries, London, Croom Helm.

Evert, S. 2005. The CQP Query Language Tutorial. Technical report. Institut fur Maschinelle Sprachverarbeitung, Universitat Stuttgart.

Kilgarriff, A. et al. 2014. The Sketch Engine: ten years on. In Lexicography (2014): 1-30.

Mcauliffe, K. 2009. Translation at the Court of Justice of the European Communities. In:

Stein D, Olsen F, Lorz A (eds) Translation Issues in Language and Law, Palgrave Macmillan.

Mcauliffe, K. 2010. Language and the Institutional Dynamics of the Court of Justice of the

European Communities: Lawyer-Linguists and the Production of a Multilingual Jurisprudence, in Gueldry M (eds) How Globalizing Professions Deal With National

Languages: Studies in Cultural Conflict and Cooperation, Lewiston, Queenstown, Lampeter: The Edwin Mellen Press.

Paunio, E. 2011. Beyond Words: The European Court of Justice and Legal Certainty in Multilingual EU Law. Helsinki, Unigrafia.

Scott, M. 2008, WordSmith Tools version 5, Liverpool: Lexical Analysis Software.

Schmid, H. 1994. Probabilistic Part - of - Speech Tagging Using Decision Trees. In Proceeding of International Conference on New Methods in Language Processing, Manchester, UK.

The moves and key phraseologies of corporate governance reports

Martin Warren
Hong Kong Polytechnic University
martin.warren@polyu.edu.hk

1 Introduction

In recent years, corporate governance has become a major issue after a number of high profile cases in which companies were found to have a lack of or inadequate corporate governance, such as the collapse of Enron in 2001 and WorldCom in 2002, and, more recently, the money-laundering scandal involving HSBC in 2012. Corporate governance covers discipline, independence, fairness, transparency, responsibility, accountability, and social awareness (Gill 2002). Good corporate governance is considered essential to improve economic efficiency, enhance the market and investors' confidence, and maintain the stability of the financial system (Ho 2003: 55). In Hong Kong, as in many countries, it is important that companies adhere to strong corporate governance since it holds the key to sustaining the growth of the city's economy, stock market, supports investor confidence, attracts international capital, and creates liquidity (Chamber of Hong Kong Listed Companies, 2007). Listed companies are required by the Stock Exchange of Hong Kong to submit interim and annual corporate governance reports, and make these reports available on their websites. While the significance of corporate governance reports is well-recognised, there is a lack of research regarding the language and discourse organisation of this genre.

2 Data

This paper offers a partial description of this relatively new genre by examining the corporate governance reports of a larger selection of major companies in Hong Kong in terms of their generic move structure and key phraseologies associated with these moves. It adopts the 'top-down' corpus-based approach to discourse analysis (Biber, Connor and Upton 2007) and combines both quantitative and qualitative approaches to discourse studies of language use. Importantly, the study is the product of collaboration between the research team, Hong Kong-based companies and professional associations. The collaboration has meant that the advice of experts in the field of corporate governance has fed into the design of the study, the data collection process, data analysis and interpretation.

The one million word Hong Kong Corpus of Corporate Governance Reports is comprised of the corporate governance reports of 217 companies listed on the Hong Kong Stock Exchange and is weighted to represent the different sectors of the Hong Kong Stock exchange (i.e. financial, property, utilities and commercial and industrial). The corporate governance reports were downloaded from the websites of listed companies in Hong Kong with the permission of the companies.

3 Findings

First, the rhetorical moves of each corporate governance report were identified to establish which were obligatory and which were optional. The move structure of the reports was then compared to the guidelines set out by the regulatory authority which include both mandatory disclosure information and recommended disclosures in order to determine the extent to which the companies simply meet or go beyond the stipulated requirements. The results of the move-structure analysis show that most of the companies simply comply with the regulations in terms of the information required to be presented in their corporate governance reports. Some of the companies included additional moves in their reports which are not required by the authorities, for example, remuneration of directors and management, business ethics, and dividends. The sequencing of the moves was also established and the most frequent patterns are described. The moves are described along with the differences in sequencing. Each of the moves found in the corporate governance reports forms its own sub-corpus which has enabled the identification of move-specific phraseologies to determine the functions and/or motivation as to why companies include the obligatory and optional information.

To carry out an analysis of the key phraseologies used, the whole corpus and the move sub-corpora were examined using ConcGram 1.0 (Greaves, 2009) to generate not only word frequency lists used to produce key words lists, but also two-word concgram lists. The concordances of the two-word concgrams were studied to determine which ones were meaningfully associated and which ones were simply chance co-occurrences to then arrive at key phraseologies for all of the corpora compiled in the study. Examples of some of the most frequent phraseologies in the main corpus and each of the move sub-corpora are described and discussed in terms of their form, patterns of co-selection and functions.

4 Implications

The findings of the study provide initial insights into the discursive practice and strategies adopted by listed companies in their corporate governance reports and the extent to which they are complying with or exceeding the requirements. The description and analysis of the patterns specific to corporate governance reports have implications for the learning and teaching of English for Specific purposes. The findings could be used to raise awareness and lead to a better understanding of this new genre. The methodology and findings might also inform other studies of specialised corpora and genre analysis in general.

Acknowledgements

The research described in this paper was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No. PolyU 5440/13H).

References

Biber, D., Connor, U and Upton, T. (eds.). 2007. *Discourse on the Move: Using Corpus Analysis to describe Discourse Structure*. Amsterdam/ Philadelphia: John Benjamins.

Chamber of Hong Kong Listed Companies. 2007. http://www.chklc.org/web/eng/index.htm, retrieved on 9 September 2009.

Gill, A. 2002. "Corporate governance in emerging markets - saints & sinners: who's got religion?" Symposium on Corporate Governance and Disclosure: The Impact of Globalisation. The School of Accountancy, The Chinese University of Hong Kong, February 2002.

Greaves, C. 2009. *ConcGram 1.0: a phraseological search engine*. Amsterdam: John Benjamins.

Ho, S.S.M. 2003. "Corporate governance in Hong Kong: Key problems and prospects", 2nd Edition, Copyright 2002, 2003. Centre for Accounting Disclosure & Corporate Governance. School of Accountancy, The Chinese University of Hong Kong.

The Text Annotation and Research Tool (TART)

Martin Weisser
Guangdong University of Foreign Studies
weissermar@gmail.com

1 Background

Research in (Corpus) Pragmatics has, for a very long time, been severely limited by a lack of suitably annotated corpora and research options. For this reason, most research in Corpus Pragmatics has generally been carried out on fixed, well-known, but non-exhaustive expressions, such as discourse markers, request structures, politeness formulae, etc. Even worse, research in traditional (theoretical) Pragmatics has generally remained limited to the discussion of invented examples and the level of individual sentences, often exaggerating the level and importance of indirect communicative strategies or issues of under-specification in quantification, as is evident from most of the chapters in popular handbooks on the topic, such as Horn and Ward (2006).

Most of the few existing spoken pragmatically annotated corpora that employ relatively generic speech-act tagsets, such as the SPICE-Ireland Corpus, the Pragmatically Annotated Switchboard Corpus, etc., have been annotated manually, which is a highly time-consuming and error-prone process. The annotations contained in them also use highly limited, and sometimes subjective, speech-act categories, such as a slightly augmented version of Searle's original distinction between *Representatives*, *Directives*, *Commissives*, *Expressives*, and *Declarations* for the former (Kirk 2013: 215-16), and a modified version of DAMSL (Dialogue Act Markup in Several Layers; Allen and Core 1997), SWBD-DAMSL (Jurafsky et al. 1997), for the latter (see Weisser 2014 and Weisser forthcoming for detailed discussions). One notable exception to this is the SPAAC Corpus (Leech and Weisser 2003), a small part of which has recently been published as the SPAADIA Corpus (Leech and Weisser 2013). This corpus already uses an enhanced and generic set of speech-act tags that has recently been augmented even further to form the DART tag set (Weisser 2010), which is used to automatically annotate dialogues using some 80+ speech-act combinations, including information about initiation-response/adjacency pairs, in the Dialogue Annotation and Research Tool (DART). The automated annotation in DART not only provides a considerably more fine-grained analysis/annotation, but also guarantees strict consistency, something that is almost impossible to achieve using manual annotation.

Thus, DART has already changed the situation regarding the dearth of suitably annotated corpora for spoken language by making it possible to analyse and annotate transcribed dialogues automatically, and on a large scale, most importantly on the level of speech acts. Furthermore, DART, by also identifying and annotating other pragmatics-relevant levels, such as syntax, semantics, semantico-pragmatics, and surface polarity, as well as through its built-in analysis features, additionally makes it possible to investigate form-function correspondences in detail. This already enables researchers in Pragmatics to carry out research on an unprecedented level, but unfortunately, so far, only for spoken data.

The development of the Text Annotation and Research Tool (TART), reported on here, represents an attempt to remedy this situation by expanding the potential of corpus-based pragmatic analysis and creation of annotated corpora to incorporate written texts as well. The main aim here is to explore ways of transferring and adapting, as far as necessary, the DART tagset for speech acts in spoken language to equally cover and exhaustively describe the various communicative functions of written language, retaining the notions of genericity and adaptability to different domains/genres already inherent in the DART approach.

2 The TART Design

In doing so, the TART design will retain the main original design features of DART, including the ability to pre-process, automatically annotate, post-process and analyse corpus data from within one and the same user-friendly research environment, and using a simple, highly readable, form of XML. The original analysis options already include the ability to concordance on features pertaining to the specific levels of annotation, conduct n-gram analyses based on linguistically meaningful units, as well as to use the results of such analyses to cyclically feed back into the improvement of the analysis methodology and existing annotations through the ability to directly access and edit the original data via hyperlinks from within the analysis results themselves. Some of these features have already been 'ported' to TART directly, and other, more fine-grained analysis options for the selection of data and linguistic criteria for analysis will be developed and integrated into TART throughout the ongoing design process. Yet other options that will allow the user to explore various features related to complexity will be integrated based on my experience in designing the Text Feature Analyser (Weisser 2007).

As far as the general architecture and customisability of linguistic resources is concerned, a similar model to the one adopted in DART, which al-

ready makes the lexica and some of the other resources editable, will be adopted, but a further aim is to improve these options by also exposing more of the underlying grammar and pragmatic inferencing model to even computationally relatively inexperienced linguists to enable them to create suitably customised large-scale corpora of pragmatically annotated written language, as well as analyse them on a variety of different levels, in a simple and efficient manner.

References

Allen, J. and Core, M. 1997. "Draft of DAMSL: Dialog Act Markup in Several Layers". Available from: ftp://ftp.cs.rochester.edu/pub/packages/dialog-annotation/manual.ps.gz .

Horn, L. and Ward, G. 2006. *The Handbook of Pragmatics*. Oxford: Blackwell. (paperback edition of 2004).

Jurafsky, D., Shriberg, E. and Biasca, D. 1997. "Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coder Manual". Available from: http://www.stanford.edu/~jurafsky/ws97/ics-tr-97-02.ps

Kirk, J. 2013. "Beyond the Structural Levels of Language: An Introduction to the SPICE-Ireland Corpus and its Uses". In Cruickshank, J. and McColl Millar, R. (eds.) 2013. *After the Storm: Papers from the Forum for Research on the Languages of Scotland and Ulster triennial meeting*, Aberdeen 2012. Aberdeen: Forum for Research on the Languages of Scotland and Ireland, 207-32.

Leech, G. and Weisser, M. (2003). "Generic Speech Act Annotation for Task-Oriented Dialogue". In Archer/Rayson/Wilson/McEnery (Eds.) *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster University: UCREL Technical Papers, vol. 16.

Weisser, M. 2014. "Speech act annotation". In Aijmer, K. & Rühlemann, C. (Eds.). *Corpus Pragmatics: a Handbook*. Cambridge: CUP.

Weisser, M. 2014. "DART – the Dialogue Annotation and Research Tool". Submitted to *Corpus Linguistics and Linguistic Theory*.

Weisser, M. 2014. "The DART Manual. Application manual to accompany the Dialogue Annotation & Research Tool". Available from http://martinweisser.org/publications/DART_manual.pdf.

Weisser, M. 2013; forthcoming 2015. "Corpora". In Barron, A., Gu, Y. and Steen, G. (Eds.). *The Routledge Handbook of Pragmatics*. London: Routledge.

Weisser, M. 2010. *Annotating Dialogue Corpora Semi-Automatically: a Corpus-Linguistic Approach to Pragmatics*. Habilitation (professorial) thesis, University of Bayreuth.

Weisser, M. 2007. "The Text Feature Analyser – a Flexible Tool for Comparing Different Levels of Text Complexity". In Schmied/Haase/Povolná (Eds.). *Complexity and Coherence: Approaches to Linguistic Research and Language Teaching*. Göttingen: Cuvillier Verlag. pp. 49-63.

Pop lyrics and language pedagogy: a corpus-linguistic approach

Valentin Werner
University of Bamberg
valentin.werner@uni-bamberg.de

Maria Lehl
Tonguesten
maria@tonguesten.com

1 Introduction

Although English pop music lyrics are part of most people's everyday life, to date, they have largely been ignored in corpus linguistic research. This can be deduced from the facts that lyrics rarely form part of standard corpora of English and that the amount of corpus-based research explicitly devoted to this register (e.g. Kreyer and Mukherjee 2007 or Werner 2012) is restricted. In addition, in applied linguistic attempts to exploit them for EFL teaching purposes they have mostly – and despite their high motivational value (see e.g. Syed 2001; Beath 2010; Israel 2013) – been sidelined to the role of "additional" or "light" material, usually found at the end of chapters, and barred from the use for the instruction of "serious" matter, such as aspects of grammar (see Murphey 1990, 1995 for notable exceptions).

We will argue that lyrics should emerge from their shadowy existence as they are a worthy subject both for corpus linguists and for practitioners in EFL for various reasons. With that goal in mind, we will address the topic of pop lyrics from three angles. First, we will present a general overview of linguistic features of lyrics and thus offer a brief stylistic analysis (in terms of locating lyrics in relation to other text types as well as on a written-spoken continuum), also considering learner-related aspects. Second, we will widen the perspective and will consider why the NLP annotation of lyrics is notoriously difficult due to some of their inherent features. It will also be discussed how these issues can be overcome. In the final section, we will address the question of how pop lyrics can be used in language teaching and learning (e.g. in terms of web applications such as Tonguesten's Rebeats[106] platform), taking advantage of the specific opportunities offered by a corpus-based approach.

2 Stylistic analysis

The data on which the analyses are based derive from purpose-built sources. One of them is the *Chart Corpus* (cf. Werner 2012), a 342,202-token corpus (1,128 songs) containing lyrics from songs that were highly successful (i.e. at least among the top five) in the UK and the US in the years 1946 to 2008.

A general quantitative comparison to other text types using the Multidimensional Analysis Tagger (Nini 2014) locates lyrics close to the category "informational interaction". This seems surprising as lyrics typically are viewed as a form of one-to-many communication and thus supposedly lack characteristic interactional features. However, when the individual dimensions of variation (following Biber 1988) are considered in detail, an ambiguous picture emerges. The analysis yields lyrics as an involved text type (Dimension 1), but with non-narrative concerns (Dimension 2), for instance. This ties in with previous research which has shown that lyrics can be viewed as a "particular" genre that (i) cannot unequivocally be assigned to the written or spoken mode and (ii) that is characterized by features associated both with formal and informal usage (Werner 2012: 43).

Learners, who receive their formal English instruction largely with the help of textbooks (which aim at a standard form of the target language) may be unfamiliar with a number of nonstandard features that occur in lyrics. Potential hurdles are contractions (*upon* > *'pon*), as well as other elisions, for instance of auxiliaries or third-person markers, all illustrated in (1).

(1) but she gone and she not comeback me beg her please 'pon me knees and she still never stop (Pato Banton: "Come back")

Another case in point are nonstandard pronoun and verb forms, as in (2) or (3), which may be used as identity markers or can also be interpreted as devices to indicate the cultural hybridity of the text (e.g. realized through a combination of standard and Creole features).

(2) so me say, we a go hear it on the stereo (Musical Youth: "Pass the Dutchy")
(3) but me know I'm not a fear to you (Sean Paul and Blu Cantrell: "Breathe")

The motivation of being able to cope with such nonstandard features as well as with the inherent hybridity of lyrics renders them a challenging but equally stimulating resource for learners. Likewise, reliable NLP annotation of lyrics – with available taggers usually trained on standard forms of a language – is a challenging task for the corpus linguist, as will be shown subsequently.

3 NLP annotation

Part-of-speech (POS) tagging is a gateway step into

[106] http://www.rebeats.tv

corpus linguistic research of lyrics. However, training a POS tagger would require the annotation of a sufficiently large training corpus that does not exist up to date. In an exploratory study, six pre-trained tagger models using the Penn tag set were assessed on a 100-song gold standard, compiled from the top ten UK albums of the years 2001 to 2011 (Lehl 2014).

With a focus on testing a broad range of tagging approaches, the HunPos tagger (Halácsy et al. 2007), the Stanford tagger (Toutanova et al. 2003) and SVMTool (Giménez and Màrquez 2004) were selected. All tagger models are trained on the Wall Street Journal (WSJ) corpus. However, some models use online chat conversations, Tweets and other web content as additional training data.

The results show tagger model performances for lyrics ranging between 90.60% and 93.05% and thus well below the state-of-the-art of 97-98% on the WSJ corpus. The best-performing model was the Stanford tagger model, which is trained on the WSJ corpus enriched by a chat corpus and Tweets.[107]

Knowing that all models are trained on the WSJ corpus among others, the taggers were assessed separately on WSJ-tokens, which had been encountered by all taggers in their WSJ training data (*known tokens*), and on non-WSJ tokens. A qualitative analysis shows that, apart from noise-related tagging errors, many of the inaccuracies on the non-WSJ tokens can be traced back to lyrics-specific phenomena, primarily contractions (see above) and musical tropes (such as *yeah* and *woah*). Most tagging errors of this type can easily be avoided by using word lists and regular expressions. However, the low accuracy of taggers on non-WSJ tokens contributes only little to the inferior general performance of taggers on lyrics as compared to the performances of taggers on the WSJ corpus. Even on the known tokens alone the maximum tagging accuracy lies at merely 93.52%. An error analysis using confusion matrices revealed common tagging errors to be standard tag confusions, such as the following:

- VB wrongly tagged as VBP

  Have/VBP* yourself a merry little Christmas"
  (Michael Bublé: "Have Yourself a Merry Little Christmas")

- VBN wrongly tagged as VBD

  Have you heard/VBD* the news today
  (P!nk: "Gone to California")

This poses the question why these common tagging errors occur more frequently in lyrics. One possible explanation is that sentence boundaries are mostly missing. As a consequence, each lyric line was fed to the taggers as one sentence, which may have given insufficient context for tagging. Other possible explanations are that lyrics contain significantly more occurrences of elisions and non-standard grammar (see above) than the training data of the taggers.

These results suggest that increasing the size of provided context for tagging (e.g. by pairwise binding of consecutive lines) and compiling a sufficiently large training corpus are necessary steps of research to engage in. However, an important point of investigation that has to be undertaken before is the computation of a ceiling tagging performance by an inter-rater agreement. A preliminary evaluation indicates that the potential for tagging ambiguity in lyrics is generally higher than in other text types. This is due to the musically imposed shortness of lines, the frequency of elliptic constructions, incoherent content, and slang, which make tagging sometimes challenging even to the human annotator.

4 Integrating corpus linguistics and language learning

The limitations illustrated should not disguise the fact that a lot can already be done with annotated lyrics data, and that it is viable to go beyond traditional uses of lyrics in educational contexts. A central field for the application of linguistically annotated lyrics corpora is represented by Computer- and Mobile-Assisted Language Learning (CALL and MALL).

There has been a recent trend of language learning gamification, one phenomenon being online language courses that target the use of lyrics and song videos for EFL. Rebeats is one example of such a web-based EFL platform in development that uses linguistically annotated data and offers one road of how findings from lyrics-related linguistic research can be applied. The main goal of the platform is to automatize the creation of language exercises from lyrics, packing the outcome into an engaging multiple-choice game as exemplified in Figure 1.

In this case, POS-tagged lyrics are used to automatically create a verb tense exercise targeting the construction of the present perfect in English. The learner is challenged by multiple choice exercises while the video clip is playing, and receives instant feedback.

Examples such as Rebeats show that an integration of corpus-based findings and application in language learning is possible. In the future, the linguistic community should provide more insights (i) on individual features of "popular" content (see

[107] https://gate.ac.uk/wiki/twitter-postagger.html

also Bértuoli-Dutra 2014), (ii) on how to deal with them in NLP, and (iii) on how to exploit their full pedagogical potential.

References

Beath, O. 2010. "'I want to be more perfect than others': a case of ESL motivation." Paper presented at the Faculty of Education and IERI HDR Conference, Wollongong, 12 November 2010. Available online at http://ro.uow.edu.au/edupapers/161/

Bértuoli-Dutra, P. 2014. "Multi-dimensional analysis of pop songs." In T. B. Sardinha and M. V. Pinto (eds.) *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber.* Amsterdam: Benjamins. 149-176.

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Giménez, J. and Màrquez, L. 2004. "SVMTool: a general POS tagger generator based on Support Vector Machines." In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa and R. Silva (eds.) *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Paris: ELRA. 43-46. Available online at http://www.lrec-conf.org/proceedings/lrec2004/pdf/597.pdf

Halácsy, P., Kornai, A. and Oravecz, C. 2007. "HunPos – an open source trigram tagger." In *45th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Interactive Poster and Demonstration Sessions.* Prague: Association for Computational Linguistics. 209-212. Available online at http://aclweb.org/anthology/P07-2

Israel, H. F. 2013. "Language learning enhanced by music and song." *Literacy Information and Computer Education Journal* 2 (1): 1269-1275.

Kreyer, R. and J. Mukherjee. 2007. "The style of pop song lyrics: a corpus-linguistic pilot study." *Anglia* 125 (1): 31-58.

Lehl, M. 2014. *Stairway to Learner's Heaven: Using Song Lyrics to Build a Resource for Automatic Creation of Language Exercises.* Unpublished Master's thesis, University of Osnabrück.

Murphey, T. 1990. *Song and Music in Language Learning: An Analysis of Pop Song Lyrics and the Use of Song and Music in Teaching English to Speakers of Other Languages*. Frankfurt: Peter Lang.

Murphey, T. 1995. *Music and Song.* Oxford: Oxford University Press.

Nini, A. 2014. *Multidimensional Analysis Tagger 1.2*. Available online at http://sites.google.com/site/multidimensionaltagger

Syed, Z. 2001. "Notions of self in foreign language learning: a qualitative analysis." In Z. Dörnyei and R. Schmidt (eds.) *Motivation and Second Language Acquisition*. Honolulu: University of Hawai'i Second Language Teaching and Curriculum Center. 127-148.

Toutanova, K., Klein, D., Manning, C. and Singer, Y. 2003. "Feature-rich part-of-speech tagging with a cyclic dependency network." In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. 173-180. Available online at http://nlp.stanford.edu/pubs/tagging.pdf

Werner, V. 2012. "Love is all around: a corpus-based study of pop music lyrics." *Corpora* 7 (1): 19-50. Available online at http://www.euppublishing.com/doi/pdfplus/10.3366/cor.2012.0016

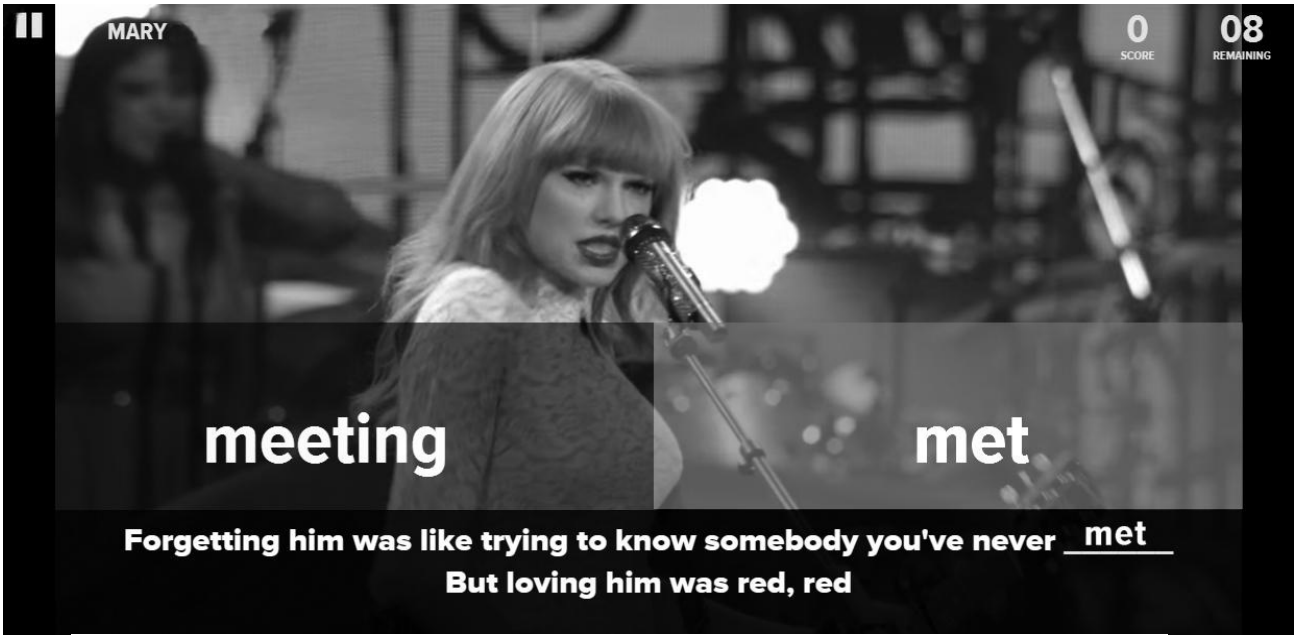


Figure 1. Example of a learning exercise on the onlkine language platform Rebeats

Multimodal resources for lexical explanations during webconferencing-supported foreign language teaching: a LEarning and TEaching Corpus investigation.

Ciara R. Wigham
Université Lumière Lyon 2
ciara.wigham@univ-lyon2.fr

1 Outline of the research question

Within the computer-assisted language learning (CALL) field, multimodal research endeavours to consider the simultaneous presence and interaction between verbal communication modes (audio, text chat) present in foreign language learning situations with co-verbal and non-verbal modes (gestures, gaze, posture, other kinesic aspects). This paper explores how trainee-teachers of French as a foreign language, during webconferencing-supported teaching, orchestrate different semiotic resources that are available to them for lexical explanations.

2 Study context and participants

The pedagogical context is a telecollaborative project where 12 trainee teachers of French as a foreign language met for online sessions in French with 18 undergraduate Business students from an Irish university. The participants met for seven 40-minute online sessions in autumn 2013 via the webconferencing platform *Visu* (Bétrancourt, *et al.,* 2011). Each online session was thematic and focused on Business French.

A research protocol was designed around this learning context. Data produced *during* the learning project itself was collected (webcam videos, text chat messages, audio recordings of collective feedback session with the trainee teachers, reflective reports), as well as data produced uniquely *for* the research project (observation notes, post-course questionnaires and interviews). Participation in the research study was voluntary - all 12 trainee teachers (ten females, two males) and 12 students (eight females, four males) gave permission to use their data.

3 Staged methodology of the LEarning and TEaching Corpus approach

The data collected have been structured into a LEarning and TEaching Corpus (Wigham *et al.,* 2014). Reffay *et al.* define a LEarning and TEaching Corpus (LETEC) "as a structured entity containing all the elements resulting from an online learning situation, whose context is described by an educational scenario and a research protocol" (2012:15). It comprises a XML "manifest" that describes the corpus' components: the learning design, the research protocol, the interaction data, all participants' productions and licences relating to ethics and access rights (see Figure 1). The XML schema allows interactions from different tools and environments to be stored and described in a standardized way, facilitating data analysis.

In the CALL field, multimodal LETEC provide resources for second language development, teacher education research and also teacher training (Wigham & Chanier, 2014). LETEC differ from *learner corpora* in that they do not comprise uniquely data from test situations not focus uniquely on learners' productions but the learning context and other course participants (tutors, native speakers...) (see Reffay *et al.,* 2008).

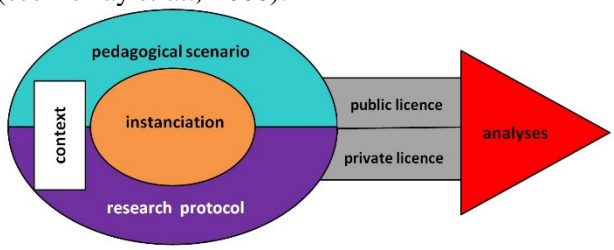


Figure 1: LETEC components

This paper will, firstly, detail the staged methodology for building a LETEC (see Figure 2), including the challenges for data collection when video recordings of the participants are concerned and how the different institutions ethical constraints were considered prior to the corpus creation and the dissemination of selected sub-sets of the corpus among the CALL research community (see also Blin *et al.,* 2014). The implications of these challenges and constraints on methodological choices will be reflected upon.

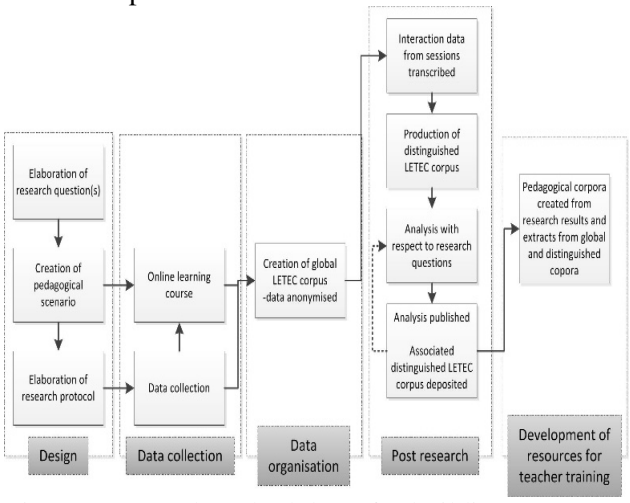


Figure 2: Staged methodology for building a LETEC

4 Lexical explanations and multimodality

In the second part of the paper, the LETEC will be investigated to show how the trainee-teachers orchestrated different semiotic resources for lexical explanations. In face-to-face contexts, multimodality helps teachers explain the nuances of lexical items, reinforce verbal messages through illustration and reduce ambiguity (Kellerman, 1992; Lazaraton, 2004).

To study lexical explanations with regards to a weconferencing teaching context, a sub-set of the corpus will be used. It comprises primarily the webcam and *hors-champ* videos of three trainee teachers engaged in interactions (see Figure 3). Audio recordings of five trainee feedback sessions and extracts of trainee post-course interviews supplement the analysis.

The webcam interactions were transcribed (see Table 1) and annotated using the software ELAN (Sloetjes & Wittenburg 2008). Transcriptions of the feedback sessions were also completed. For the trainee post-course interviews, we proceeded by a global exploration phase of the audio recordings that allowed remarks and comments pertinent to our research question to be identified.

The paper will report quantitatively on the number of lexical explanations given in the webconferencing sessions examined and report on the different communication modes and modalities utilised. Then, in order to 'zoom in' on fine-grained detail, a qualitative analysis will examine several lexical explanation episodes to show how trainee teachers coordinated different communication modalities simultaneously to facilitate their lexical explanations.

An example of this fine-grained analysis is illustrated in Figure 3. The trainee-teacher combines audio and kinesics modalities to explain lexical item 'volunteer' (*bénévole*, in French): she combines the audio modality with a culturally specific emblem in the kinesics modality to illustrate 'earning money', then a self-deictic gesture to accompany the phrase 'I'm a volunteer' before using an abstract deictic gesture moving back and forth between the students' and trainee teacher's communication space to illustrate the difference in their situations. The corpus demonstrates how different multimodal resources are mobilized during lexical explanations. The trainee feedback session data and trainee interview data will complement the analysis by showing the importance that the trainee teachers attributed to the multimodal nature of the webconferencing environment.

Mode	*Modality*	*Act type*	*Explanation*
Verbal	Audio	Audio act	Verbal act in the full duplex audio channel
		Silence	Interval between two audio acts greater than three seconds
	Text chat	Text chat act	Message entered into the text chat window
Co-verbal	Kinesics	Communicative gestures	Gestures seen in the webcam recordings (*iconic, metaphoric, deictic, beat, emblem, communicative action*)
	Kinesics	Mimics	Facial expressions seen in the webcam recordings and their functions (e.g. surprise, happiness, incomprehension)
Non-verbal	Kinesics	Extra-communicative gestures	For example, scratching forehead, pushing hair behind ear, 'playing' with pen.

Table 1: Multimodal transcription categories

Figure 3: Orchestration of multimodal resources during the lexical explanation of '*bénévole*' with webcam and *hors-champ* views shown

5 Perspectives

To fully understand the contribution of multimodality to webconferencing-supported teaching, both the teacher's and learners' contributions to the interaction must be studied. This paper paves way for further analyses of the corpus that examine how the trainee teachers' lexical explanations were received by the learners. The

interest of organising data into LETEC in which the pedagogical design and research protocol are described is seen here: the corpus can be examined by researchers not originally involved in the pedagogical project for cumulative analyses.

Acknowledgements

This research was supported by the Ulysses programme funded jointly by the Irish Research Councils and the French Ministry of Foreign Affairs.

References

Bétrancourt, M., Guichon, N. & Prié, Y. (2011). Assessing the use of a Trace-Based Synchronous Tool for distant language tutoring. Proceedings of the 9th International Conference on Computer-Supported Collaborative Learning, Hong-Kong, July 2011. pp.478-485

Blin, F., Guichon, N., Thouësny, S. & Wigham, C.R. (2014). Creating and sharing a language learning and teaching corpus of multimodal interactions: ethical challenges and methodological implications. Sixteenth International CALL Research Conference, 7-9 July, Antwerp, Belgium.

Kellerman, S. (1992). 'I see what you mean': The Role of Kinesic Behaviour in Listening and Implications for Foreign and Second Language Learning, Applied Linguistics, 13(3). pp.239-258.

Lazaraton, A. (2004). Gesture and speech in the vocabulary explanations of one ESL teacher. A microanalystic inquiry, Language Learning, 54 (1). pp.79-117.

Reffay, C., Betbeder, M-L. & Chanier, T. (2012). Multimodal learning and teaching corpora exchange: lessons learned in five years by the Mulce project. International Journal of Technology Enhanced Learning, 4(1). pp.11-30.

Reffay, C., Chanier, T., Noras, M. & Betbeder, M-L. (2008). Contribution à la structuration de corpus d'apprentissage pour un meilleur partage en recherche. Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation (Sticef), 15. [oai: edutice.archives-ouvertes.fr:edutice-00159733].

Sloetjes, H. & Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

Wigham, C.R. & Chanier, T. (2014). Pedagogical corpora as a means to reuse research data and analyses in teacher-training. In Colpaert, J., Aerts, A. & Oberhofer, M. (Eds). Research Challenges in CALL. Proceedings of the Sixteenth International CALL Conference, 7-9 July Antwerp: University of Antwerp.

Wigham, C.R., Thouësny, S., Blin, F. & Guichon, N., (2014). ISMAEL LEarning and Teaching Corpus. Dublin, Ireland: Centre for Translation and Textual Studies & Lyon, France: Laboratoire Interactions, Corpus, Apprentissages & Représentations.

Size isn't everything: Rediscovering the individual in corpus-based forensic authorship attribution

David Wright
Nottingham Trent University
david.wright@ntu.ac.uk

1 Forensic authorship attribution

Forensic authorship attribution is concerned with identifying the author(s) of disputed, questioned or anonymised documents that are potentially evidential in alleged infringements of the law or threats to security. In forensic casework, the linguist is tasked with identifying stylistic consistencies and differences between these 'disputed' texts and texts which are 'known' to have been written by the suspect(s) involved. Unfortunately, analysis is often difficult given that disputed texts in such casework—whether they are emails, text messages, tweets, letters or suicide notes (among other text types)—can be 'unhelpfully short' (Coulthard and Johnson 2007: 172). At the same time, the known texts, provided by the police or legal teams and used for comparison by the linguist, usually comprise 'any old collection of texts' (Cotterill 2010: 578).

2 Trends in authorship research

As a result of these practical challenges, empirical authorship attribution research, most of it computational or 'stylometric' in nature, has aimed to test how effective their proposed methodologies are when attributing small amounts of data (e.g. Eder 2013; Luyckx and Daelemans 2011; Rico-Sulayes 2011; Grant 2007). Such studies invariably return results which show that as the size of the datasets used is systematically reduced, the accuracy of their approach decreases. Alongside these studies, focus has also been on which kinds of linguistic features are most useful for attributing texts to their correct author, ranging from frequency of function words to syntactic part-of-speech clusters (e.g. Koppel et al. 2009; Grieve 2007; Chaski 2001).

Although the value of such research is clear, the pre-occupation with overall corpus size and the sets of linguistic features used has resulted in the neglect of analysing closely the individual authors who make up the corpora. Authorship attribution is practiced on the assumption that each person has their own distinctive idiolect (Coulthard 2004: 431). By extension, it may also be reasonable to assume that different authors' idiolects are manifest in different ways. Therefore, while the use of function words may be useful for capturing distinctive aspects of one author's style, for example, it may be useless for another author. Yet, while authorship analysts accept the existence of idiolect, there is no similar acknowledgement of the fact that different authors' idiolects may be identifiable in different ways. Consequently, poor results for a particular methodology lead to generalised conclusions that either the linguistic features used are ineffective, or the method is not robust to decreases in dataset size, without considering how effective the method was for the individual authors within the data.

3 Research questions

This research develops a methodology for authorship attribution which uses word n-grams (strings of *n* words) to correctly identify anonymised email samples. The analysis is driven by three related research questions:

- How does reducing the amount of data to be attributed affect the success of the method?
- How successful is the methodology for the different authors who make up the corpus?
- Which length of n-gram is most effective in identifying authors?

4 The Enron Email Corpus

Enron is a former energy trading company based in Houston, Texas. As part of the legal enquiry into the company's controversial bankruptcy, a dataset of employees' emails was made publicly available. The data is available in various versions online, but the source used for this study is that provided by Carnegie Mellon University (CMU) (Cohen 2009).

After the CMU set had been 'cleaned up' and prepared for use in authorship research, the final corpus used includes emails of 176 Enron employees, totalling over 60,000 emails and 2.5 million tokens. This particular study focuses on a sub-corpus of twelve of the 176 authors, which contains 12,633 emails and 382,070 tokens, and is labelled the 'Enron Email Corpus 12-author sample' (EEC12).

5 Method

This study uses word n-grams to attribute authorship. Corpus linguistic and psycholinguistic research has argued that the associations which people make between words and their subsequent production of collocation patterns are unique to individuals (Barlow 2013; Hoey 2005; Wray 2002; Nattinger and DeCarrico 1992; Sinclair 1991). Therefore, word n-grams of between one and six words in length were drawn upon here to capture such idiolectal and author-distinctive collocation and co-selection patterns.

The attribution experiment itself involved

extracting ten different random email samples of 2%, 5%, 10%, 15% and 20% from the sets of each of the EEC12 authors. The resulting 600 samples represent the 'disputed' texts, and range in size from 55 to 14,859 tokens. Using a bespoke computer program called *Jangle* (Woolls 2013), each of these individual samples was compared against (i) the remaining emails of the author from whom the sample was taken and (ii) the full email sets of the other 175 Enron employees. These sets represent the 'known' data. The program measured how similar the 'disputed' samples were to the 'known' email sets in terms of how many one to six word n-grams they shared, and used Jaccard's co-efficient (Juola 2013; Grant 2013) to produce a similarity statistic for each comparison. If the author to whom the sample belonged obtained the highest Jaccard score of all then the attribution was successful. If another of the 176 candidate authors achieved the highest score, then attribution was unsuccessful.

6 Results and discussion

In total there were 3,600 pair-wise tests in the experiment: 600 samples being attributed by six different word n-gram lengths. The results of these tests can be used to answer the three research questions stated above.

First, reducing the amount of data drastically affects the accuracy of the method. The average success rate for attribution of 20% samples, ranging from 48 to 459 emails and 762 to 14,859 tokens in size, was 92.6% across the six n-gram lengths. By the time the samples had been reduced to 2% of the authors' sets, ranging from 4 to 45 emails and 55 to 950 tokens, the average accuracy rate had declined to as low as 17.1%.

Second, the method was far more successful when attributing the samples of some authors than others. Given the seemingly clear relationship between success rate and sample size, one would expect the method to work best with those authors who have the largest 'disputed' sample sizes (in terms of tokens). However, this is not the case. The email samples of John Lavorato, a former president of Enron, were the easiest to attribute of all EEC12 authors, with the method achieving a success rate of 80.7% with his samples. In terms of size, Lavorato has only the fifth largest samples of the EEC12 authors. Similarly, Jim Derrick, Enron's chief lawyer, has the smallest sample sizes in EEC12. Despite this, he ranks as eighth easiest of twelve to identify using this method. The four authors with whom the method performs worse than Derrick all have considerably larger 'disputed' samples than him. In fact, there is little to no relationship between how large an author's samples compared with the other authors and how successful the method is in correctly identifying their writing styles. Results of this kind show that word n-grams are better at capturing the idiolect of some authors than they are for others. In turn, this indicates that some authors' idiolects are manifest in the distinctive collocation and co-selection choices they make, while for other authors this seems to be less true.

Finally, the most successful n-gram length overall was four-grams, correctly identifying the authors of samples in 70.7% of the tests in which they were used. This suggests that strings of four consecutive words are most effective in capturing distinctive elements of idiolect. Again, however, the results varied for the different EEC12 authors. Four-grams only outperformed the other five measures for four of the twelve authors; trigrams, five-grams or six-grams performed best across the remaining eight. This provides evidence to suggest that not only are word n-grams generally better at identifying some authors than others, but specific lengths of n-gram capture specific idiolects more effectively.

7 Implications

The results of this study reveal that while corpus size is important, it is not everything in forensic authorship attribution. The method employed here does generally perform far better on larger datasets. However, a closer examination of this general result found that it worked well for some authors and not others, regardless of how comparatively large their disputed samples were.

By focusing attention on comparing the effectiveness of different linguistic features and how robust methods are to reductions in dataset size, authorship research has lost sight of the linguistic individual. In the future, rather than hastily judging the quality of a method on the basis of overall results, forensic linguists should consider how successful their methods are in identifying individual authors and idiolects *within* these corpora. This is, after all, the purpose of authorship attribution.

References

Barlow, Michael. 2013. Exemplar theory and patterns of production. Paper presented at *Corpus Linguistics 2013*, Lancaster, 22–26 July 2013.

Chaski, Carole, E. 2001. Empirical evaluations of language-based author identification techniques. *Forensic Linguistics: (The International Journal of Speech Language and the Law)* 8(1), 1–65.

Cohen, William W. 2009. *Enron Email Dataset*. [online]. Available from: http://www.cs.cmu.edu/~enron/. [Accessed November 2010].

Cotterill, Janet. 2010. How to use corpus linguistics in forensic linguistics. In Anne O'Keefe and Michael

McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 578–590.

Coulthard, Malcolm. 2004. Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics* 24(4), 431–447.

Coulthard, Malcolm and Alison Johnson. 2007. *An Introduction to Forensic Linguistics: Language in Evidence*. London: Routledge.

Eder, Maciej. 2013. Does size matter? Authorship attribution, small samples, big problem. *Literary and Linguistic Computing* [online]. Available from: http://llc.oxfordjournals.org/content/early/2013/11/14/llc.fqt066.full [Accessed June 2014].

Grant, Tim. 2007. Quantifying evidence in forensic authorship analysis. *International Journal of Speech Language and the Law* 14(1), 1–25.

Grant, Tim. 2013. Txt 4N6: Method, consistency and distinctiveness in the analysis of SMS text messages. *Journal of Law and Policy* 21(2), 467–494.

Grieve, Jack. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing* 22(3), 251–270.

Hoey, Michael. 2005. *Lexical Priming: A new theory of words and language*. London: Routledge.

Juola, Patrick. 2013. Stylometry and immigration: A case study. *Journal of Law and Policy* 21(2), 287–298.

Koppel, Moshe, Jonathan Schler and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science And Technology* 60(1), 9–26.

Luyckx, Kim and Daelemans, Walter. 2011. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing* 26(1), 35–55.

Nattinger, James R. and Jeanette DeCarrico. 1992. *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Rico-Sulayes, Antonio. 2011 Statistical authorship attribution of Mexican drug trafficking online forum posts. *The International Journal of Speech, Language and the Law* 18(1), 53–74.

Sinclair, John. M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Woolls, David. 2013. *CFL Jaccard n-gram Lexical Evaluator* (*Jangle*) version 2. CFL Software Limited.

Wray, Alison. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Illuminating President Obama's Argumentation for Sustaining the Status Quo, 2009 – 2016

Rachel Wyman
King's College London
rachel.wyman@kcl.ac.uk

1 Introduction

In 2008 a financial crisis caused by the financial elite exacerbated the already staggeringly unequal wealth distribution of the United States, further shifting resources from the bottom 90% to the top 10%. In the midst of the ensuing economic recession, Barrack Obama was elected President. Vowing to fight government corruption, the discourse of his campaign speeches aligned with the middle and working classes, promising a return to a fair system based on unity. Yet in 2010, Obama used taxpayer money to bail out the country's most powerful banks, shattering any hope of true change. In doing so, the President chose to support a hegemonic system where the rich are given a separate set of rules.

Obama is feted for his rhetorical ability. But critical appreciation of his argumentation for supporting the aforementioned hegemony is sorely lacking. This presentation will detail a methodology for analyzing this based on combining corpus linguistic and argumentation analysis in order to investigate Obama's argumentation, examining his 'rhetoric of equality' and comparing this to the reality of what his policies have produced.

This paper presents a new approach to analyzing political texts based on argumentation analysis. It utilizes the argument reconstruction framework from *Political Discourse Analysis* (Fairclough & Fairclough 2012) and works to extend use of this framework through an innovative combination of qualitative software analysis (NVIVO) and corpus linguistic analysis (WMatrix). Fairclough and Fairclough (2012)'s analytical framework is based around five functional categories (see below). I use NVIVO to code these functional argument units in Obama's speeches, spanning 2009 – 2016. This is followed by a corpus linguistic analysis of these distinct functional unit corpora. The goal is to examine in detail how different functional argument units work linguistically to construct Obama's presidential rhetoric, which in turn sustains the aforementioned hegemony.

2 Political Discourse Analysis (2012)

In Political Discourse Analysis (2012) Norman and Isabela Fairclough present a framework for analyzing arguments functionally. The framework enables deliberation on an argument put forward by a politician. Fairclough and Fairclough's framework is valuable to the extent that it avoids thinking about argumentation monolithically and thus misleadingly. The framework captures five functions of political argument:

- 1. *Claim for action*: Agent ought to do *A*.
- 2. *Goal*: Agent's goal is a future state of affairs in which Agent's value commitments are realized.
- 3. *Circumstances*: Agent's context of action – natural, social and institutional facts.
- 4. *Values*: Agent's concerns and value commitments.
- 5. *Means-goal*: Action A is the means that will take Agent from *C* to *G* in accordance with *V*.

. Fig. 1 shows how F and F combine these functions.

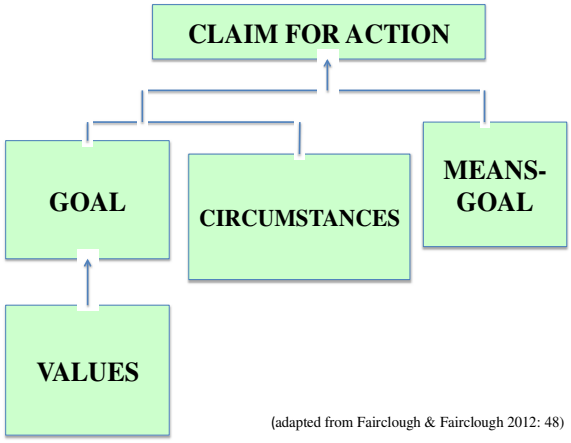


Figure 1

The way in which the speaker represents the nation's current circumstances and values enters into his/her claim. However, these representations are not always accurate. False representations may enter into discourse, resulting in flawed narratives based on arguments that do not stand up to critical evaluation. Yet they still form the premises for arguments about how to respond to political problems with action. For example, Obama's argument for Wall Street reform depends on his depiction of Wall Street as responsible for causing the 2008 financial crisis. Wall Street reform will result in the return to a fair system. However, the U.S. government regulates the financial system; therefore it is implicated in its reckless behavior. Yet Obama's narrative succeeded and the Wall Street Reform Bill was passed, demonstrating how language can be used to corrupt the political process by creating a reality that is inaccurate.

3 Research questions

- How have Obama's recurrent arguments in support of action been constructed in his Weekly Addresses over his two terms? Are there patterns and regularities across them?
- Do these arguments stand up to critical evaluation? Are the representations of reality that construct Obama's depictions of our current circumstances and values, and which enter into his claims for action, accurate?
- How do Obama's extra-discursive actions as president compare to his narrative?

3.1 Corpus Analysis

A corpus analysis of Obama's 400 speeches over his two terms will show overall patterns, identifying the narrative being constructed. Who is depicted as being the obstacle preventing us from attaining equality? A reference corpus of the Weekly Addresses from the presidencies of Ronald Reagan, Bill Clinton and George W. Bush will be used for comparison.

3.2 Coding of Functional Units

Using NVIVO to code the individual Obama speeches will illuminate the arguments that he makes over the course of eight years. These arguments can be broken down into separate functional units: *claim for action, goal, means-goal, circumstances* and *values.*

3.3 Corpora of Functional Units

By creating separate corpora for the functional units, the patterns in these corpora and the relationship between them can then be examined. What are Obama's *claims for action*, *goals* and *means-goals?* How does he connect them? How does he represent the country's *circumstances* and American *values*? How do these representations feed into his arguments? How do Obama's smaller arguments connect to form major arguments?

3.4 Argument Analysis

The fourth step is to analyze these arguments, demonstrating where they deconstruct, and looking at how representations enter into the discourse and the argument. By incorporating the corpus data it is possible to analyze the arguments he makes within individual speeches, but also the overarching arguments that these smaller ones construct. How

does Obama describe our world in order to eliminate options that conflict with his agenda?

3.5 Rhetoric vs. Reality

The final part of this project involves looking at Obama's executive orders and what legislation he has supported, signed into law and vetoed. Is his narrative in support of equality supported by the reality of his actions? If not, who do his arguments and actions ultimately support?

This methodology can be used to effectively identify how the main topics of Obama's speeches connect and form the arguments that come to define presidential discourse over significant periods of time.

Translation as an activity of under-specification through the semantic lenses

Jiajin Xu
Beijing Foreign Studies University
xujiajin@bfsu.edu.cn

Maocheng Liang
Beijing Foreign Studies University
liangmaocheng@bfsu.edu.cn

1 Corpus-based studies of Translation Universals

The hypotheses of Translation Universals (TUs) have been influential since Baker (1993), but not without challenges. The conceptualisation of TUs is innovative in the sense that translation is seen as a legitimate variety of language, variously known as 'translationese', 'interlanguage', 'hybrid language', 'the third code', etc. The popularity of TUs related research has practically been enabled by automatic analysis of lexical and syntactic features in translated texts. A great deal of research has tried to test one or more of the hypotheses, namely, characteristic textual features of translation such as explication, disambiguation or simplification, normalisation, etc.

2 What current corpus-based TUs studies fail to capture?

Previous corpus based TUs studies have focused on the over- and under-representation of surface lexical and grammatical features. For instance, how certain word forms (e.g. hapax legomena, abnormal collocations, connectives, foreign words, type-token ratio, etc.) and/or word classes (e.g. pronouns, nouns, content words, function words, or the ratios based on these) occur in translated texts. Additionally, some length measures (e.g. mean word/sentence length) have been taken into account as well. Typically, the linguistic features are compared with those used in the target or original texts.

Little however has been done in terms of the semantic aspects of the features of translation, because, as it is commonly known, the semantic dimension of language is not amenable to automatic analysis. Within the overarching concept of semantics, lexico-semantics sees the most exciting computational advances in natural language processing, for instance, the development of the lexical database WordNet, MRC psycholinguistic database, The Edinburgh Associative Thesaurus (EAT), the UCREL Semantic Annotation System (USAS), and latent semantic analysis (LSA). In the

present study, WordNet and MRC lexico-semantic resources are exploited to measure the lexical specificity or depth of translated and original English texts. EAT has been integrated into the latest version of MRC. USAS serves a similar purpose as WordNet does in this case. Latent semantic analysis concerns more of the lexico-semantics at the textual level, which is slightly beyond the scope of the present study.

3 Lexico-semantic features for specificity

Word specificity in texts is examined from the following eight lexico-semantic aspects based on WordNet and MRC resources.

A) Polysemy for content words;
B) Hypernymy for nouns;
C) Hypernymy for verbs;
D) Hypernymy for nouns and verbs;
E) Familiarity for content words;
F) Concreteness for content words;
G) Imagability for content words; and
H) Meaningfulness for content words;

Greater values for A) through D) indicate that more general words are used. Content words (category A) cover broad range of words, which subsumes category D and can be further broken down into category B and C. Categories E to H originated from psycholinguistic norming tests with native English speakers. The specificity is often regarded as cognitive-based indices (Crossley 2008), because general and specific words require mentally different effort to comprehend the texts.

4 The experiment

4.1 Comparable corpora

The datasets used for the present study include 88,177 words of translated English editorials and review texts from Chinese source texts, i.e. the Marco Polo corpus, and 90,312 words of original English editorials and review texts, i.e. the Crown B-C corpus. The translated English texts were collected from the 'The Marco Polo Project' web site (http://marcopoloproject.org) which is non-profit-making project initiated by a group of Chinese-culture loving Australians and some overseas Chinese. All English translations on this website are produced by users on a volunteer basis. The original English texts come from the text categories B and C of 2009 Brown family type of English corpus (Xu and Liang, 2013) featuring editorials, reviews and commentaries on books, movies, and social issues. The translated texts gathered from the Marco Polo project web site share very similar themes to the Crown B-C corpus. All the texts in both corpora were published or translated in the last five years. So the two corpora are comparable in terms of both size, date of production, and most importantly content. They form the empirical basis of the present small-scale comparative study of the linguistic features of English translation from Chinese.

4.2 Lexico-semantic features retrieval

The present study collected the values of eight lexico-semantic features of both the original and the translated English texts using the Coh-Metrix online tool 3.0 (http://tool.cohmetrix.com). The eight measures are part of the Coh-Metrix 3.0 measure set containing 106 linguistic features.

4.3 Comparison of the lexico-semantic features in the corpora

Independent Samples T test was used to gauge the difference between Marco Polo and Crown B-C texts along the eight lexico-semantic dimensions.

Categories	Marco Polo	Crown B-C	t score	p value
A: Polysemy	3.77	3.75	.56	.57
B: HyperNoun	6.25	6.17	.89	.38
C: HyperVerb	1.58	1.70	-6.25	.00
D: HyperNV	1.71	1.97	-8.53	.00
E: Familiarity	571.32	563.89	7.242	.00
F: Concreteness	370.27	378.32	-2.79	.01
G: Imagability	405.50	411.83	-2.41	.02
H: Meaningfulness	430.51	428.15	1.37	.17

Table 1: Lexico-semantic comparisons between Marco Polo corpus and Crown B-C corpus

The results show that statistically significant differences are found in lexico-semantic categories C) HyperVerb, D) HyperNV, E) Familiarity, F) Concreteness and G) Imagability. The difference of WordNet measures C and D shows that the original English texts (Crown B-C) demonstrate greater word specificity than do the translated English texts (Marco Polo), as hypernymy (i.e. lexical specificity) in WordNet locates words on a hierarchical scale allowing for the measurement of the number of subordinate words below and super-ordinate words above the target words (McNamara, et al. 2014: 76). The significant difference of hypernymy for both nouns and verbs (cf. HyperNV) is identified, and interestingly the hypernymy of verbs contributes to the difference while hypernymy of nouns does not. This last point merits further exploration into the actual verbs used in the texts.

The under-representation of concreteness and imagability in translated texts seems to corroborate the WordNet based study of lexical specificity from psycholinguistic evidence. In other words, from a reader's perspective, translated English texts are less concrete and less likely to construct mental images.

The over-representation of lexical familiarity actually reiterates the same lexico-semantic

judgement, because familiar words are correlated with high-frequency abstract words. Greater value of familiarity shows more vague and abstract expression.

To sum up, this small-scale comparable English corpora based study takes into account the hypernymic knowledge, or lexical depth, of both translated English and original English texts and revisits the TUs research on the basis of surface lexical and grammatical features.

Concluding remarks

As is commonly acknowledged in the translation community, translation is an activity of interpreting of the meaning of a text in the source text and reproducing it in another language. Hence, the study of semantic aspects, or linguistic meaning, of translated texts is implicit in the endeavour of TUs research in the first place. Features of translation, however, are by no means a monolithic thing and should be triangulated from lexical, syntactic, textual and semantic aspects.

References

Baker, M. 1993. "Corpus linguistics and translation studies: Implications and applications". In M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and technology: In honour of John Sinclair*. 233-250. Amsterdam: John Benjamins.

Crossley, S., Greenfield, J., and McNamara, D. 2008. "Assessing text readability using cognitively based indices". *TESOL Quarterly* 42 (3): 475-493.

Coltheart, M. 1996. MRC psycholinguistic database: Machine usable dictionary. http://www.psych.rl.ac.uk/MRC_Psych_Db_files/mrc2.html (accessed on 9 Jan. 2015).

McNamara, D., Graesser, A., Philip M. McCarthy, P., and Cai, Z. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

Xu, Jiajin and Liang, Maocheng. 2013. "A tale of two C's: Comparing English varieties with Crown and CLOB (The 2009 Brown family corpora)". *ICAME Journal* 37: 175-1

Construction of a Chinese learner corpus: Methods and techniques

Hai Xu
Guangdong University of Foreign Studies
xuhai1101@gdufs.edu.cn

Richard Xiao
Lancaster University
r.xiao@lancaster.ac.uk

Vaclav Brezina
Lancaster University
v.brezina@lancaster.ac.uk

1 Introduction

Most of the learner corpora that have been developed so far, as Pravec's survey (2002) shows, are confined to English learner corpora, such as ICLE, JPU, ICANE, CLEC and SWECCL. However, with the rise of China as a global power, Chinese as a major world language has become an increasingly popular foreign language. The number of L2 Chinese learners is growing rapidly. In teaching and learning Chinese as a foreign language, a learner corpus plays an instrumental role.

2 Publically available Chinese learner corpora

Two Chinese learner corpora are publically available: HSK 动态作文语料库 (Monitoring Corpus of HSK Essays) [108] and 留学生书面语与口语语料库 (Overseas Students Written and Spoken Chinese Corpora)[109].

The Monitoring Corpus of HSK Essays, which consists of 4.24 million of tokens, was developed by a team at Beijing Language and Culture University. It covers only one specific type of written data: written essays by HSK[110] takers from 101 countries between 1992 and 2005. Out of the 11,569 pieces of essays in it, the corpus is skewed towards learners from East and Southeast Asian countries: Korean (4,171), Japanese (3,211), Singapore (843), Indonesian (739), Malaysian (422), Thai (374), Vietnamese (227), and Burmese (202). The corpus data have been annotated at the levels of Chinese characters, phrases, sentence patterns, etc., but the annotation scheme is very complicated. The online version provides the concordance function, but no functions of wordlist, *n*-gram and keyness.

The Overseas Students Written and Spoken

[108] http://202.112.195.192:8060/hsk/login.asp

[109] Written Corpus: http://www.globalhuayu.com/corpus3/Search.aspx ; Spoken Corpus: http://www.globalhuayu.com/corpus5/Default.aspx

[110] HSK is a Chinese proficiency public test, similar to IELTS or TOEFL.

Chinese Corpora were built by a team at College of Chinese Language and Culture, Jinan University in Guangzhou. With the size of the around 4 million of tokens, the written corpus contains the examination essays as well as free compositions by L2 Chinese learners from 47 countries between 2001 and 2010. The written corpus has not been annotated yet, and the online version is also limited to the concordance function. The size and data types of the spoken corpus are unknown, and the data were claimed to be produced by L2 Chinese learners from 22 countries. The data do not contain any transcription conventions and error annotation, and only limited search function is provided.

3 Guangwai-Lancaster Chinese Learner Corpus

As the discussion above suggests, very few Chinese learner corpora have been constructed. Even those publically available, they have collected only a specific type of data, and are biased towards Korean, Japanese and Southeast Asian speakers. Thus, a balanced Chinese learner corpus is called for.

With a competitive grant from British Academy IPM 2013 Scheme, a team at Lancaster University and Guangdong University of Foreign Studies in Guanghzou has been developing a balanced Chinese learner corpus. The corpus, which is called 'Guangwai-Lancaster Chinese Learner corpus', contains around 1 million tokens of both written and spoken data.

While designing the new corpus, we have tried to meet nearly all the criteria discussed in Granger (2013), including the variables pertaining to the task (medium, topic, and timing), and to the learner (proficiency level, mother tongue background, and gender).

We adjusted the original ratio of written vs. spoken data from 7:3 to 6:4, for we found more spoken data are available, and more importantly, it reflects the predominance of speech in daily communication.

The corpus covers a variety of task types. The written corpus data range from essays under test condition to free compositions. And the spoken corpus data cover utterances in oral tests and free conversations. The subtypes of oral tests data include conversations between one or two (or three) L2 speakers and an examiner, and monologues by an L2 examinee. As for free conversations, there are monologues on either the topic 'My Hometown' or "A Memorable Trip". The corpus also contains conversations between a native speaker (i.e. a postgraduate student majoring in teaching Chinese as a foreign language) and a nonnative speaker (i.e. an L2 learner).

In terms of the learner's proficiency level, we originally set the ratio of beginner vs. intermediate vs. advanced to 3:4:3, but we have to adjust the ratio to approximately 2:5:3 for written data, and around 4:5:1 for spoken data. In practice, we observed that L2 Chinese beginners can produce more spoken data than written data. And as a rule, while learning a foreign language, there are a larger number of L2 beginners, and fewer learners can reach the advanced level. In addition, the syllabus for the advanced learner is focused more on written forms than on speaking.

In terms of the learner's L1 background, we have collected written corpus data produced by L2 learners from 64 countries, and spoken data from 72 countries.

The ration between the male and female speaker in the corpus is around 45% to 55%, as we have originally planned.

In the process of transcription and digitalization, we need to retain some transcription conventions, so that they can represent some features of an L2 speaker's writing and speaking. Thus, some information like meta data will not be missing. For spoken corpus data, they include speakers ID, fillers (like 呃, 嗯, 啊, 哦, 噢), short or long pauses, meta-linguistic behaviour and comments (such as <clear_throat>, <sniffle>, <laugh>, <sigh>, <sneeze>, <whisper>), foreign language, and name and number anonymisation. For written corpus data, we mainly transcribed typos. Different from the Western language, the typos with Chinese characters include additional or missing strokes, and a nonexistent character coined by a speaker

Last but not least, the corpus contains over 40 learners' longitudinal corpus data, which we can 'track the same learners over a particular period [i.e. at different proficiency level]' (Granger 2013).

Acknowledgements

This project was supported by a grant from British Academy International Partnership & Mobility 2013 Scheme.

References

Granger, S. 2013. "Learner corpora" In C.A. Chapelle *The encyclopedia of applied linguistics*. London: Blackwell.

Pravec, N.A. 2002. "Survey of learner corpora". *ICAME Journal* 26: 81-114.

Automatic Pattern Extraction: A Study Based on Clustering of Concordances

Tao Yu
Beijing Foreign Studies University
yutaowy@163.com

1 Background

Language patterns are ubiquitous in running texts. Traditional methods to language pattern recognition often involve reading or sampling concordances with "colored-pen method" (Kilgarriff and Koeling 2003). However, it is time-consuming and labor-intensive. Even worse, untypical patterns, quite often, are submerged in heterogeneous data (Sinclair 2003). Furthermore, automatic pattern recognition is under way (Mason 2004; Mason and Hunston 2004) since "we look forward to the development of an automatic pattern identifier" (Hunston and Francis 2000: 272); however, previous studies don't show promising results.

To tackle the fore-mentioned problems, this study aims to automatically retrieve verb patterns by simulating manual work of reading concordances, mainly based on the clustering of concordance lines with the aid of similarity measure and KMeans algorithm.

2 This Study

This study combines rule-based and statistics-based methods to extract verb patterns from automatically grouped concordance lines. Theoretically, this study is based on Pattern Grammar (ibid) and Verb Pattern List summarized by Francis et al. (1996), and statistically employs similarity measure (Euclidean distance) and KMeans algorithm.

The five-step procedures are as follows: 1) Extract and pos-tag concordance lines or sentences containing the node word; 2) Closely examine Verb Pattern List, then summarize the necessary elements in verb patterns, finally build feature sets; element features in verb patterns covering different linguistic levels, such as words, word combinations, word classes, syntactic and semantic features. 3) Transform linguistic data in concordances into codes in the feature sets; 4) Assign feature weights and position weights to features in the transformed concordances, then build feature-concordance matrix. Measure similarity between each two columns (each column stands for one concordance), then based on the similarity scores, employ KMeans algorithm to cluster concordances into groups. 5) Extract common features in each grouped transformed concordances, then combine the node and common features according to their original order, then verb pattern in each group of concordances is automatically extracted. Finally, verb pattern list of the node is generated.

All the above five steps are interlinked with each other; especially the later four procedures are the core of automatic pattern extraction model. If there is anything wrong with any step, the output will not satisfactorily meet the aim of the study.

To fulfill the above tasks, Perl is utilized as the programming language, and four Perl modules are involved, namely, Transform.pm, Feature.pm, Kmeans.pm and LCS.pm. Transform.pm is used to fulfill the task of transforming linguistic data in concordances into codes in the feature sets. Feature.pm is utilized to assign feature weights and position weights to features in the transformed concordances, then build feature-concordance matrix. Kmeans.pm is called to calculate the similarities among concordances and group the concordances. LCS.pm is to extract common features in each group of concordances.

Furthermore, seven configuration files are utilized when calling the module Transform.pm, just name a few, word.txt, word_comb.txt, and location_weight.txt. What's more, some parameters should be readjusted for different nodes or different number of concordances, when executing KMeans.pm.

3 Result and discussion

To test the validity of the model, concordances with manual pattern labels (testing set) are classified into groups twice. Firstly, the number of concordance groups (K) is set according to manual classification. Secondly, K is set based on the internal validity measure (Residual sum of squares) of KMeans.

This study yields the following results: 1) Distributions of patterns and pattern elements are different from each other among the verbs. Top 5 patterns in each verb pattern list are different and verbs show different tendency to co-occur with different features. 2) The selection of K based on internal validity measure yields better results than pre-set K. Meantime, the model is much more flexible, since potential users can choose different Ks according to different purposes, based on internal validity measure. 3) Comparing the results of fore-mentioned twice clusterings of concordances, average precisions of automatic pattern extraction are 90.99% and 95.91% respectively, with an increase of 9.99 and 14.91 percentage rates than the average precision of automatic verb pattern recognition mentioned in previous studies.

The above-mentioned findings verify some hypothesis, such as, distribution hypothesis, the inseparability of lexis and grammar.

4 Applications

This model designed in this study can be applied to a wide range of studies, such as language pedagogy, and language studies.

This model will be of great help to the design of lexical and grammar syllabus, still will be very valuable in language teaching, such as provision of typical language patterns and typical concordances.

Language studies will benefit a lot from this model. For example, it can be applied to dictionary compilation, genre analysis, translation studies and contrastive interlanguage analysis.

References

Francis Gill, Susan Hunston and Elizabeth Manning 1996. *Collins Cobuild Grammar Patterns 1: Verbs.* London: HarperCollins.

Hunston, Susan and Gill, Francis (2000). *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English.* Amsterdam: John Benjamins.

Kilgarriff, Adam and Rob Koeling 2003. An evaluation of a lexicographer's workbench incorporating word sense disambiguation. *Proc. CICLING, 3rd Int Conf on Intelligent Text Processing and Computational Linguistics,* Mexico City. Springer Verlag.

Mason, Oliver 2004. Automatic Processing of Local Grammar Patterns. *In Proceedings of CLUK.* University of Birmingham, 166–171.

Mason, Oliver and Susan, Hunston 2004. The automatic recognition of verb patterns--A feasibility study. *International Journal of Corpus Linguistics* 9(2):253-270.

Sinclair, John 2003. *Reading Concordance.* London: Pearson Education Ltd., Longman.

Exploring the variation in World Learner Englishes: A multidimensional analysis of L2 written corpora

Yu Yuan
University of Leeds
mlyy@leeds.ac.uk

Selinker(1971) proposed that interlanguage characterizes as systematic and dynamic throughout the stages of second language acquisition. In other words, the interlanguage which the learner has constructed is portrayed as an internally consistent system and the process of development from one stage to the next is ordered and regular (Ellis 1985:118). It has become increasingly acknowledged, however, that interlanguage is also variable. It is believed that the variety in a learner's language can be a part of his learning process as well as contextual alterations. Systematicity and variability are two reconcilable features of learner language.

Contextual variability, as the second type of variability identified in interlanguage, is evident when the language user varies his use of linguistic forms according to the linguistic environment. Then, a full account of the situational and contextual variability in interlanguage requires studying how the linguistic environment constrains the operation of interlanguage rules at different stages of development in different contexts. Interlanguage language of learner productions in various societies, therefore as linguistic system in its own right, offers a valuable resource of studying how they are varying systematically due to the linguistic, situational and psycholinguistic factors that are imposed on the learners in different ethnic groups. It is worth attempting this with the advent of large corpora of learner written texts and the exponentially increasing computing power. Hundt and Mukherjee (2011:2) argued that it is high time learner Englishes and second-language varieties are described and compared on an empirical basis in order to draw conceptual and theoretical conclusions with regard to their form, function and acquisition. They believe such descriptive studies and comparisons were not possible on a large scale 20 years back as the relevant ESL (e.g. the International Corpus of English, ICE) and EFL (e.g. the International Corpus of Learner Corpus, ICLE) computerized corpora have only become available in recent decades. With this purpose in mind, the researcher compares corpora of written English texts by Chinese English learners and learners from 11 other countries (i.e.,

the major components of ICLE 1.1) to describe the features of EFL of different learner varieties in its entirety so that the hidden patterns of systematic variances of using linguistic forms for different ethnic learner groups can be to some extent uncovered.

In retrospect, comparative analyses of learner Englishes are largely based on certain individual features, which cannot give a full picture of systematic variation across varieties. For example, Carolin Biewer (2011) only analyzes the use of modal auxiliaries across the comparable corpora; Gaëtanelle Gilquin and Sylviane Granger (2011) investigate the use of proposition into in 4 components of ICLE and compare it with native British English; Benedikt Szmrecsanyi and Bernd Kortmann (2011) analyze and compare the degrees of grammatical analyticity and grammatical syntheticity across a wide range of components of ICE and ICLE. In order to compare learner varieties on a more macro-level, it necessitates a new comparative approach that enables comparisons across varieties as a whole. Multi-dimensional analysis, a corpus-based research approach developed for the comprehensive analysis of register variation, can be utilized to achieve the goal of this research: to explore the variation between learner English varieties. Biber (2009:823) posits that many registers are distinguished only by a particularly frequent or infrequent occurrence of a set of register features. This quantitative comparative approach allows us to treat each learner English variety "as a continuous construct: texts are situated within a continuous space of linguistic variation" and enables us to analyze how learner Englishes are " more or less different with respect to the full range of core linguistic features" (Biber 2009:824). This approach is advantageous in that it can circumvent the problem of considering individually too many linguistic characteristics and their idiosyncratic distributions and base the analyses on the co-occurrences and alternation patterns for groups of linguistic features, thus important differences across learner English varieties are likely to be revealed.

This study is based on the 67 linguistic features (as in Biber 1988) and 14 syntactic features selected by Lu (2010). Such an enhanced model of MD analysis with syntactic complexity metrics on the one hand lends insight into the developmental variation in terms of syntactic complexity in EFL writings by learners at different countries or regions, and on the other hand supplements the tendency of over-emphasis of lexical features in the original model. Linguistic features either positively or negatively loaded on certain factors (hidden patterns of co-occurrences) for specific learner varieties will be reported as systematic variation on a micro-level scale. Through the factor analysis of 12 sub-corpora, the researcher identified 4 dimensions (principal component analysis and varimax rotation was used in R Pscych package), the Kaiser-Meyer-Olkin measure of sampling adequacy was .723, above the recommended value of .6, and Bartlett's test of sphericity was significant (χ^2 (3916) = 298402.183, $p < .05$): syntactic complexity vs. simplicity (which explains 9% of the variance), sentence builder methods (no negative loadings, which explains 8% of the variance), elaboration vs. judgment (which explains 7% of the variance) and interactive vs. informative discourse (which explains 6% of the variance) A significance test of variances is then used to tell if there any meaningful difference among these learner English varieties on each factor extracted. For example, the results show that Italian variety is most prominent on the second dimension, while Czech learner English is most representative of the first dimension and Spanish Learner English is close to both dimension 1 and dimension 2. As Xiao (2009) compared world Englishes along different registers (e.g. argumentative essays and literature examination paper per ICLE) factor by factor, register-based comparison of learner Englishes will also be reported in the study.

References

Biewer, C. (2011). Modal auxiliaries in second language varieties of English: A learner's perspective. In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins Publishing, pp. 7–33.

Ellis, R. (1985). Sources of variability in interlanguage. Applied Linguistics, 6(2):118-31.

Gilquin, G. & Granger, S. (2011). From EFL to ESL: Evidence from the International Corpus of Learner English. In J. Mukherjee & M. Hundt (eds). *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins Publishing, pp.55-78.

Szmrecsanyi, B. & Kortmann, B. (2011). Typological profiling: Learner Englishes versus indigenized L2 varieties of English. In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins Publishing, pp. 167–187.

Biber, D. (1988). Variation across Speech and Writing. Cambridge: Cambridge University Press.

Biber, D. (2009). Theory-driven corpus research: using corpora to inform aspect theory. In A. Lüdeling & M.Kyto (eds) *Corpus Linguistics: An International Handbook [Volume 2]*. Berlin: Mouton de Gruyter. 823-855.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474-496. [L2 Syntactic Complexity Analyzer].

Mukherjee, J. & Hundt, M. (2011). *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins Publishing.

Selinker L. (1972). Interlanguage. International Review of *Applied Linguistics,* 10, 209-241.

Xiao, Z (2009) Multidimensional analysis and the study of world Englishes. *World English* , 28(4): 421-450.

Nativeness or expertise: Native and non-native novice writers' use of formulaic sequences

Nicole Ziegler
University of Hawai'i at Manoa
nziegler@hawaii.edu

Formulaic language has been shown to be an integral part of native and non-native language use, with research demonstrating the importance of learners' use of formulaic language as a measure of second language (L2) development (Ellis 1996; Ellis and Simpson-Vlach 2008). In addition, recent research has demonstrated that multi-word units, such as formulaic sequences and lexical bundles, are distributed differently depending on register (Biber 2006; Biber and Conrad 1999; Biber et al. 2004), discipline (Hyland 2008a) and proficiency or writing skills (Chen and Baker 2010; Cortes 2002, 2004; Hyland 2008b; Staples et al. 2013). Findings also suggest that novice native speaker academic writers often do not use the same sequences found in expert academic writing, with results indicating significant differences in functional use, frequency, and variation across genres and proficiency level (Hyland 2008).

Other studies have investigated the distribution of multi-word combinations between published academic prose and first (L1) and second language (L2) student academic writing. For instance, Chen and Baker (2010), using small L1 and L2 English sub-corpora from the British Academic Written English (BAWE) corpus, compared the use of four-word bundles in novice student writing with the academic prose found in the Freiburg-Lancaster-Oslo/Bergen (FLOB) corpus. Distributional and functional analyses indicated that published writing had more referential expressions and fewer discourse markers than were found in the student writing (Chen and Baker 2010). Römer (2009) found that there were few differences in the phraseological profile across native and advanced non-native undergraduate writers, suggesting that learners' proficiency in academic writing, rather than nativeness, may play a substantial role in the use of formulaic language by both L1 and L2 writers. However, although these findings indicate interesting similarities in the frequency and length of formulaic sequences by native and non-native writers, further research is needed to examine whether nativeness plays a role in writers' functional use of formulaic language.

Overall, much of the research involving lexical bundles and multi-word sequences has focused on register variation (e.g. Biber et al. 2004) or

comparative investigations between novice and professional native speaker writing (e.g. Hyland 2008) or native and advanced non-native writers (e.g. Ädel and Erman 2012). In an effort to move beyond the distinction of performance according to native or non-native speaker status (Swales 2004), the current study seeks to extend previous research (Chen and Baker 2010; Römer 2009) and gain further insight into the similarities and differences between native and learner novice academic writing by examining the following research questions:

1. What are the most frequent four-word lexical bundles in novice unpublished L1 and L2 academic writing?
2. What are the differences between the use of lexical bundles in novice unpublished L1 and L2 academic writing?

Quantitative and qualitative analyses were used to examine the differences between the frequency and functional use of four-word formulaic sequences in L1 and L2 novice academic prose. Using the British Academic Written English (BAWE) corpus and the International Corpus of Learner English (ICLE), sub-corpora of L1 and L2 academic writing were created, and n-gram analysis was used to identify the four-word clusters and their frequencies.

Following previous research, the cut-off frequency of four-word formulaic sequences was set at 10 times per million words (Biber et al. 2004). To protect against idiosyncratic use of individual writers, sequences must have occurred in a minimum of five different texts in order to be included in the analyses. In addition, overlapping lexical bundles that occurred the same number of times, such as *can be seen in* and *as can be seen,* thus suggesting these two four-word bundles were constituents of a five-word bundle, were removed to prevent inflated results. Sequences were analyzed for functional variation and were categorized as referential expressions, stance, or discourse organizing sequences, following the discourse function classification proposed by Biber and colleagues (Biber et al. 2004; Biber and Barbieri 2007).

Although type and token analyses revealed similarities between the L1 and L2 corpora, preliminary results indicated that the distribution and frequency of use differed across L1 and L2 novice writers, with the L1 corpus demonstrating more balanced use and more even distribution of four-word combinations. In other words, despite the number and occurrences of bundles displaying similar frequency within the two corpora, L2 writers had a much more narrow distribution of bundles, with the most common bundle occurring more than twice as often as the next most frequent bundle. This uneven frequency rate suggests that L2 writers may have a greater reliance on the most common bundles, underscoring the need for improved explicit instruction and awareness raising of lexical bundles in L2 classrooms. Analyses also suggest additional differences between L1 and L2 writers in the types of formulaic sequences across both function and structure, with L1 writers demonstrating more similarities to expert native academic texts.

Overall, findings support previous research addressing the similarities between L1 and L2 student writing (Chen and Baker 2010; Römer 2009), and suggest that although there may be differences across native and non-native writers, certain features of novice academic writing may be attributable to writing or education levels rather than nativeness. Although differences were found across L1 and L2 student writing, the fact that novice writers, regardless of native language, seem to have the same gaps in academic writing expertise suggests that both populations would benefit from more explicit training on how to produce more proficient academic writing. In addition, direct instruction and awareness raising may improve learners' distributional use of sequences, helping learners to use a wider range of formulaic language in academic writing, as well as demonstrate the range of functions formulaic sequences can serve in academic prose.

5 Acknowledgements

Some of the data in this study come from the British Academic Written English (BAWE) corpus, which was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800).

References

Ädel, A. and Erman, B. 2012. "Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach." *English for Specific Purposes 31:* 81-92.

Biber, D. and Barbieri, F. 2007. "Lexical bundles in university spoken and written registers." *English for Specific Purposes 26: 263-286.*

Biber, D., and Conrad, S. 1999. "Lexical bundles in conversation and academic prose." In H. Hasselgard and S. Oksefjell (eds.), *Out of Corpora. Studies in honour of Stig Johansson*. Amsterdam: Rodopi.

Biber, D., Conrad, S., and Cortes, V. 2004. "If you look at…Lexical bundles in university teaching and

textbooks." *Applied Linguistics 25*: 371-405.

Chen, Y., and Baker, P. 2010. "Lexical bundles in L1 and L2 academic writing." *Language learning and technology 14:* 30-49.

Cortes, V. 2002. "Lexical bundles in freshman composition." In R. Reppen, Fitzmaurice, S. M., and Biber, D. (eds.), *Using corpora to explore linguistic variation*. Amsterdam: John Benjamins.

Cortes, V. 2004. "Lexical bundles in published and student disciplinary writing: Examples from history and biology." *English for Specific Purposes 23:* 397-423.

Ellis, N. 1996. "Sequencing in SLA: Phonological memory, chunking, and points of order." *Studies in Second Language Acquisition 18:* 91-126.

Ellis, N. and Simpson-Vlach, R. 2008. "Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL." *Teachers of English to Speakers of Other Languages Quarterly 42: 375-396.*

Hyland, K. 2008a. "As can be seen: Lexical bundles and disciplinary variation." *English for Specific Purposes 27*: 4-21.

Hyland, K. 2008b. "Academic clustering: Text patterning in published and postgraduate writing." *International Journal of Applied Linguistics 18:* 41-62.

Römer, U. 2009. "English in academia: Does nativeness matter?" *Anglistik: International Journal of English Studies 20:* 89-100.

Staples, S., Egbert, J., Biber, D., and McClair, A. 2013. "Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section." *Journal of English for Academic Purposes 12:* 214-225.

Posters

The development of an Arabic corpus-informed list of formulaic sequences for language pedagogy

Ayman Ahmad Alghamdi
University of Leeds

1 Summary

This study aims to construct an Arabic corpus-informed list of formulaic sequences for language pedagogy. The selection of formulaic sequences in this list will be based on several pedagogically-relevant criteria from the perspective of second language comprehension (e.g high frequency and meaningfulness). Further, two studies will be conducted in order to demonstrate how this list should be presented. These studies will exemplify the type of applications in language research and pedagogy in which the list may be beneficially employed. The first project targets to integrating the Arabic phrasal expressions list with the design of a teaching material for Arabic as a second language learners while the second one aims to develop a test of the Arabic formulaic sequences. The pedagogical implications of this list are estimated to promote Arabic teacher and researchers' awareness of the imperative need for the inclusion of formulaic sequences in the process of teaching and learning Arabic language to non-native speakers.

2 Background and Rationale to the Study

The last three decades have seen a marked increase in pedagogical interest in formulaic language phenomenon (e. g. lrujo, 1986; Pawley and Syder, 1983; Sinclair, 1987; Wray, 2002). Pawley& Syder's (1983) point out the importance of remembered phrases in the development of native-like competence of speech. Kjellmer (1990) also addresses the central role of formulaic sequences in second language learning, he stresses that collocation mostly features the nature of native language. He believes that most speech and writing of second language learners unacceptable to native ears because they mostly have a few collocation which lead them to create a new structures that might be seems strange to native speakers (1990, pp. 123-124). Therefor according to Kjellmer the focus in teaching and learning of foreign languages should be on the collocations in which they normally occur instead of individual words (1990, p. 125). Cortes (2004) asserts that Formulaic sequences promotes natural and proficient language use

He states that the "use of collocations and fixed expressions has been considered a marker of proficient language use" (p. 398) Psychological models of automaticity in language processing provide a strong support for the claim that Formulaic language promotes fluency (e.g. De Keyser, 2001; Schmidt, 1992;Segalowitz, 2003; Segalowitz & Hulstijn, 2005). These studies demonstrate the important role of formulaicity in the successful language processing.

In English language, Studies of formulaic language leading to new approaches of comprehending language, new theories of language processing and acquisition which is ultimately result in the development of new methods of teaching and learning English as a foreign language. Examples of these studies can be observed through the development of several list of English multiword expressions that can be used to help inform such instruments of L2 pedagogy as language textbooks and language tests. (e.g, Leech et at, 2001, Simpson-Vlach & Ellis, 2010, Martinez,2011)

These research paves the way for many pedagogical implications on teaching and learning English a s a foreign language.

It should be mentioned that the field of teaching Arabic to non-native speakers in SA has also witnessed some developments over the past few years. These developments have taken various forms involving the development of teaching methods, the formation of new syllabuses and the funding of several research on teaching and learning Arabic language. Nevertheless, formulaic language and the role of multiword expression in the process of learning and teaching Arabic language have not received adequate attention from either teachers or researchers. While a considerable amount of research has taken place concerning the development of formulaic sequences lists in a Teaching of English to Speakers of Other Languages (TESOL) context, to the best of my knowledge, no list of Arabic formulaic sequences has been attempted in a Teaching of Arabic to Speakers of Other Languages (TASOL) context. This study, therefore, aims to contribute to the remedying of this deficiency by constructing a frequency informed and pedagogically relevant list of Arabic formulaic sequences.

The conduct of this study will addresses the imperative need for formulaic language research particularly, within TASOL context. It noteworthy that much of the work on compiling teaching material and the development of language tests in Arabic is still based on existing lists of single orthographic words which developed a long time ago by using a very old method.so, this list estimated to be used as a pedagogical interment for researchers, teachers and curriculum designers to develop new theories and methods of teaching

Arabic language to non-native speakers. In addition the development of such a list anticipated to be as an inspiration for interested researchers to develop range of more specialized phrases list for learning Arabic for special purposes.

3 Methodology

The development of this list will be based on the integration between what Nesselhauf (2004) has called the 'frequency-based approach' and 'phraseological approach' therefor, the selection of formulaic sequences in this list will be based on several pedagogically-relevant criteria from the perspective of second language comprehension (e.g high frequency and meaningfulness). The adopted corpus in this study is the 700 million words International Corpus of Arabic developed by King Abdulaziz City for Science and Technology (KACST) which involve a collection of written Modern Standard Arabic selected from a wide range of sources which is considered to represent a wide cross-section of regional variety of Arabic.

With regard to the two studies which exemplify the pedagogical use of the phrasal expressions list, the first project concerning with the design of a teaching material based on the multiword expressions list, while the second project related to the development of an Arabic phrases test.

4 Research Questions

RQ1: From the perspective of L2 comprehension, which type of formulaic sequence should be given priority?

RQ2: How can sequences of the type defined in RQ1 then be identified and put into a list?

RQ3: How many items should a list of pedagogically-relevant formulaic sequences contain?

RQ4: How should a pedagogically-relevant list of formulaic sequences be presented?

References

Arnon, I. and Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1), 67-82.

Barfield, A. and Gyllstad, H. (2009). *Researching collocations in another language: Multiple Interpretations*. Basingstoke: Pelgrave Macmillan.

Bishop, H. (2004). Noticing formulaic sequences: A problem of measuring the subjective. LSO *Working Papers in Linguistics*, 4,15-19.

Conklin, K. and Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers *Applied Linguistics* 29(1), 72-89.

Conzett, J. (2000). Integrating collocation into a reading and writing course. In: M. Lewis (Ed.), *Teaching collocation: Further developments in the Lexical Approach* (70-86). Hove: LTP.

Corrigan, E. A. Moravcsik, H. Ouali, and K. M. Wheatley (Eds.), *Formulaic language volume 2: Acquisition, loss, psychological reality, and functional explanations* (323-346). Amsterdam: John Benjamins Publishing Company.

Cowie, A. P. (1992). Multiword lexical units and communicative language teaching. In P. J. L. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp.1-12). Houndsmills: Macmillan.

Cowie, A. P. (1998). *Phraseology: Theory, analysis, and application*s. Oxford: Oxford University Press.

Coxhead, A. J. (2000). A new academic word list. *TESOL Quarterly*, 34,213-238

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, 28(3), 157-169.

Durrant, Philip Lee (2008) High frequency collocations and second language learning. PhD thesis, University of Nottingham.

Ellis, N. C. (2008). Phraseology: The periphery and the heart of language. In: F. Meunier and S. Granger (Eds), *Phraseology in language learning and teaching* (1-13). Amsterdam: John Benjamins.

Ellis, N. C., Simpson-Vlach, R. and Maynard, C. (2008). Formulaic language in native and second- language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, 42(3), 375-396.

Erman, B. (2009). *Formulaic language from a learner perspective: What the learner needs to know. In R.*

Fulcher, G. and Davidson, F. (2007). *Language testing and assessment.* New York: Routledge.

Gilner, L. (2011). *A primer on the General Service List. Reading in a Foreign Language*, 23(1), 65-83.

Grabe, W. (2004). Research on teaching reading. *Annual Review of Applied Linguistics*, 24,44-69.

Granger, S. and Meunier, F. (2008). *Phraseology: An interdisciplinary perspective*. Amsterdam: John

Howarth, P. (1998a). The phraseology of learners' academic writing. In Cowie, A. (Ed.), *Phraseology: Theory, analysis and applications* (161-186). Oxford: Oxford University Press

Hsu, J. T. (2008). Role of the multiword lexical units in currect EFL/ESL textbooks. *US-China Foreign Language*, 27-39.

Jiang, N. and Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *Modern Language Journal*, 91(3), 433-445.

Koprowski, M. (2005). Investigating the usefulness of lexical phrases In contemporary coursebooks. *ELT Journal*, 59(4), 322-332.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English based on the British National Corpus*. Harlow: Longman

Lewis, M. (2000). *Teaching collocation*. Hove: Language Teaching Publications.

Martinez, Ron (2011) The development of a corpus-informed list of formulaic sequences for language pedagogy. PhD thesis, University of Nottingham.

McEnery, T. & Hardie, A. (2012) *Corpus linguistics :method, theory and practice.* Cambridge: Cambridge University Press. 294 p.

McEnery, T. & Xiao, R. (2010) What corpora can offer in language teaching and learning *Handbook of Research in Second Language Teaching and Learning*. Hinkel, E. (ed.). London & New York: Routledge, Vol. 2, p. 364-380 17 p.

Meunier, F. and Granger, S. (Eds.) (2008). *Phraseology in foreign language learning and teaching.* Amsterdam: John Benjamins Publishing Company.

Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford: Oxford University Press.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, 5,12-25.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nesselhauf, N. (2004). What are collocations? In D. J. Allerton, N. Nesselhauf, and P. Skandera(Eds.), Phraseological units: Basic concepts and their application (1-21). Basel: Schwabe.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins Publishing Company.

Ohlrogge, A. (2009). Formulaic expressions in intermediate EFL writing assessment. In R. Corrigan, E. A. Moravcsik, H. Ouali, and K. M. Wheatley (Eds.), *Formulaic language volume 2: Acquisition, loss, psychological reality, and functional explanations* (375-386). Amsterdam: John Benjamins Publishing Company.

O'Sullivan, B. and Weir, C. J. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing: Theories and practices* (13-32). Basingstoke: Pelgrave Macmillan.

Read, J. and Nation, P. (2004). Measurement of formulaic sequences. In Schmitt, N. (Ed.), *Formulaic sequences* (23-36). Amsterdam: John Benjamins Publishing Company.

Schmitt, N. (2004). *Formulaic sequences*. Amsterdam: John Benjamins.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manua*l. Basingstoke: Palgrave Macmillan.

Schmitt, N. and Carter, R. (2004). Formulaic sequences in action: An introduction. In Schmitt, N.(Ed.) *Formulaic Sequences* (1-22). Amsterdam: John Benjamins.

Simpson-Vlach, R. and Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31,487-512.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stubbs, M., & Barth, 1. (2003). Using recurrent phrases as text-type descriminators: A quantitative method and some findings. *Functions of Language*, 10(1), 61-104.

Weinert, R. (1995). The role of formulaic language in second language acquisition: A review. *Applied Linguistics*, 16(2), 180-205.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. New York: Palgrave Macmillan.

West, M. (1953). *A general service list of English words*. London: Longman, Green and Co.

West, M. (1954). Vocabulary selection and the minimum adequate vocabulary. *ELT Journal*, VIII(4), 121- 126.

Wilkins, D. (1976). *Notional syllabuses: A taxonomy and its relevance to foreign language curriculum development*. Oxford: Oxford University Press.

Wray, A. (2000). Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics*, 21(4), 463-489.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.

Wray, A. (2009). Identifying formulaic language: Persistent challenges and new opportunities. In R. Corrigan, E. A. Moravcsik, H. Ouali, and K. M. Wheatley (Eds.), *Formulaic language volume 1: Distribution and historical change* (27-51). Amsterdam: John Benjamins Publishing Company.

Xue, G. and Nation, I. S. P. (1984). A University Word List. *Language Learning and Communication*, 3, 215-230

A Review of Semantic Search Methods to Retrieve Information from the Qur'an Corpus

Mohammad Alqahtani
University of Leeds
scmmal@leeds.ac.uk

Eric Atwell
University of Leeds
E.S.Atwell@leeds.ac.uk

The Holy Qur'an is the most important resource for the Islamic sciences and the Arabic language (Iqbal et al., 2013). Muslims believe that the Qur'an is a revelation from Allah that was given 1,356 years ago. The Qur'an contains about 80,000 words divided into 114 chapters (Atwell et al., 2011). A chapter consists of a varying number of verses. This holy book contains information on diverse topics, such as life and the history of humanity and scientific knowledge (Alrehaili and Atwell, 2014). Corpus linguistics methods can be applied to study the lexical patterns in the Qur'an; for example, the Qur'an is one of the corpora available on the SketchEngine website. Qur'an researchers may want to go beyond word patterns to search for specific concepts and information. As a result, many Qur'anic search applications have been built to facilitate the retrieval of information from the Qur'an. Examples of these web applications are Qurany (Abbas, 2009), Qur'an Explorer (Explorer, 2005), Tanzil (Zarrabi-Zadeh, 2007), Qur'anic Arabic corpus (Dukes, 2013), and Quran.com.

The techniques used to retrieve information from the Qur'an can be classified into two types: semantic-based and keyword-based. Semantic-based search techniques are concept-based which retrieves results by matching the contextual meaning of terms as they appear in a user's query, whereas the keyword-based search technique returns results according to the letters in the word(s) of a query (Sudeepthi et al., 2012). The majority of Qur'anic search tools employ the keyword search technique.

The existing Qur'anic semantic search techniques include the ontology-based technique (concepts) (Yauri et al., 2013), the synonyms-set technique (Shoaib et al., 2009), and the cross language information retrieval (CLIR) technique (Yunus et al., 2010). The ontology-based technique searches for the concept(s) matching a user's query and then returns the verses related to these concept(s). The synonyms-set method produces all synonyms of the query word using WordNet and then returns all Qur'anic verses that contain words matching any synonyms of the query word. Cross language information retrieval (CLIR) translates the words of an input query into another language and then retrieves verses that contain words matching the translated words.

On the other hand, keyword-based techniques include keyword matching, the morphologically-based technique (Al Gharaibeh et al., 2011), and use of a Chabot (Abu Shawar and Atwell, 2004). The keyword matching method returns verses that contain any of the query words. The morphologically-based technique uses stems of query words to search in the Qur'an corpus. In other words, this technique generates all other forms of the query words and then finds all Qur'anic verses matching those word forms. The Chabot selects the most important words such as nouns or verbs from a user query and then returns the Qur'anic verses that contain any words matching the selected words.

There are several deficiencies with the Qur'anic verses (Aya'at) retrieved for a query using the existing keyword search technique. These problems include the following: some irrelevant verses are retrieved, some relevant verses are not retrieved, or the sequence of retrieved verses is not in the right order (Shoaib et al., 2009). Misunderstanding the exact meaning of input words forming a query and neglecting some theories of information retrieval contribute significantly to limitations in the keyword-based technique (Raza et al.). Additionally, Qur'anic keyword search tools use limited Islamic resources related to the Qur'an. This affects the accuracy of the retrieved results.

Moreover, current Qur'anic semantic search techniques have limitations in retrieved results. The main causes of these limitations include the following: semantic search tools use one source of Qur'anic ontology that does not cover all concepts in the Holy Qur'an, and Qur'anic ontologies are not aligned to each other, leading to inaccurate and uncomprehensive resources for Qur'anic ontology.
To overcome the limitations in both semantic and keyword search techniques, we designed a framework for a new semantic search tool called the Qur'anic Semantic Search Tool (QSST). This search tool aims to employ both text-based and semantic search techniques. QSST aligns the existing Quranic ontologies to reduce the ambiguity in the search results.

QSST can be divided into four components: a natural language analyser (NLA), a semantic search model (SSM), a keywords search model (KSM), and a scoring and ranking model (SRM). NLA tokenizes a user's query and then applies different natural language processing techniques to the tokenized query. These techniques are the following: spelling correction, stop word removal, stemming, and part of speech tagging (POS). After that, the NLA uses WordNet to generate synonyms for the reformatted

query words and sends these synonyms to the SSM and the KSM. The SSM searches in the Qur'anic Ontology database to find the related concepts of the normalised query and then returns results. At the same time, KSM retrieves results based on words matching the input words. SRM refines the results retrieved from both KSM and SSM by eliminating the redundant verses. Next, SRM ranks and scores the refined results. Finally, SRM presents the results to the user.

References

Abbas, N. H. 2009. *Quran 'search for a concept' tool and website*. MRes thesis, University of Leeds.

Abu Shawar, B. and Atwell, E. 2004. An Arabic chatbot giving answers from the Qur'an. Proceedings of TALN**. 4**(2), pp.197-202.

Al Gharaibeh, A. et al. 2011. The usage of formal methods in Quran search system. In: Proceedings of international conference on information and communication systems, Ibrid, Jordan. pp.22-24.

Alrehaili, S. M. and Atwell, E. 2014. Computational ontologies for semantic tagging of the Quran: A survey of past approaches. In: LREC 2014 Proceedings.

Atwell, E. et al. 2011. An artificial intelligence approach to Arabic and Islamic content on the internet. In: Proceedings of NITS 3rd National Information Technology Symposium.

Dukes, K. 2013. *Statistical parsing by machine learning from a classical Arabic treebank*. PhD thesis.

Explorer, Q. 2005. *Quran Explorer* [Online]. [Accessed 26 October 2014]. Available from: http://www.quranexplorer.com/Search/Default.aspx

Iqbal, R. et al. 2013. An experience of developing Quran ontology with contextual information support. *Multicultural Education & Technology Journal*. **7**, pp.333-343.

Raza, S.A. et al. An essential framework for concept based evolutionary Quranic search engine (CEQSE).

Shoaib, M. et al. 2009. Relational WordNet model for semantic search in Holy Quran. Emerging Technologies, 2009. ICET 2009. International Conference on, 2009. IEEE, 29-34.

Sudeepthi, G. et al. 2012. A survey on semantic web search engine. *International Journal of Computer Science*, 9.

Yauri, A. R. et al. 2013. Quranic verse extraction based on concepts using OWL-DL ontology. *Research Journal of Applied Sciences Engineering and Technology*. 6, pp.4492-4498.

Yunus, M. et al. 2010. Semantic query for Quran documents results. Open Systems (ICOS), 2010 IEEE Conference on, 2010. IEEE, 1-5.

Zarrabi-Zadeh, H. 2007. *Tanzil*. http://tanzil.net/

A contrastive analysis of Spanish-Arabic hedges and boosters use in persuasive academic writing

Anastasiia Andrusenko
Universitat Politècnica de València

Research on metadiscourse has been conducted since the 1980s and differences in the metadiscourse use across genres and languages have been identified in the most of the research (Crismore, 1989; Fuertes-Olivera et al., 2001; Hu & Cao, 2011; Hyland, 1998, 1999; Le, 2004; Milne, 2003). It is argued by the authors on metadiscourse that the use of metadiscourse signals the writer's involvement in the text. Hyland and Tse (2004: 156) assume that "writers use 'metadiscourse' to explicitly organize their texts, engage readers, and signal their attitudes to both their material and their audience". Depending on the purpose and the audience the writers/speakers use certain metadiscourse resources. In research articles, being the main means of academic communication, education, and knowledge creation, metadiscourse contributes to a writer's voice which balances confidence and circumspection, facilitates collegial respect, and seeks to locate propositions in the concerns and interests of the discipline (Hyland, 2005: 112).

In this study we examine the use of hedges and boosters as a category of interactional metadiscourse strategies in the genre of academic article from a comparative perspective. Hedges are linguistic means used to express uncertainty about the truth in communication. Hyland argues, that "hedging enables writers to express a perspective on their statements, to present unproven claims with caution, and to enter into a dialogue with their audiences". Boosters, on the other hand, are devices that increase certainty or conviction about the propositional content (Holmes, 1984). Although metadiscourse studies are increasingly concerned among the scholars, cross-cultural research did not receive adequate attention. A plethora of studies in metadiscourse use English as a common point of reference reflecting the importance of English as a lingua franca in the global education and research community (Markkanen et al., 1993; Mauranen, 1993; Valero-Garcés, 1996; Moreno, 1997, 2004; Mur-Dueñas, 2011).

Anthologies on contrastive rhetoric have not included studies of Spanish (Connor, 1996: 52). However, extensive research on English-Spanish contrasts has been conducted by various Spanish linguists (Dafouz-Milne, 2008; Milne, 2003, 2006; Moreno, 1997, 2004; Mur-Dueñas, 2011; Valero-Garcés, 1996).

Concerning English-Arabic contrastive studies very interesting seems the study of El-Seidi (El-Seidi, 2000). She investigated the use of validity markers and attitude markers in English and Arabic argumentative writing, comparing the use of these two categories of metadiscourse in native English and native Arabic students' argumentative essays. She observed that whereas the frequency and the preferred forms of metadiscourse categories vary, both between the native English and native Arabic sets and across L1-L2 texts of each language, these categories largely appear in the same contexts to involve the writers into texts, indicating the degree of commitment to the text and their attitude towards it.

Abdelmoneim (2009) explored interpersonal metadiscourse categories in two Egyptian newspapers concerning the 2007 "Constitutional Amendments".

Abbas (2011) investigated the similarities and differences between English and Arabic in relation to interactive and interactional metadiscourse markers in linguistics research articles (RAs), comparing 70 discussion sections of RAs in both languages. He observed that metadiscourse markers play a very significant role in linguistics RAs in both English and Arabic. His findings, however, indicate the tendency among Arab writers to exaggerated use of metadiscourse markers.

The lack of literature comparing Spanish and Arabic rhetorical conventions has been the motivation for this study. In particular, Fernando Trujillo Saéz (Sáez, 2000) underlined the need to investigate larger corpora in contrastive rhetoric and to compare Spanish with other languages, which is what this research work aims at.

The study seeks to answer the following research questions:

- Are there any differences/similarities in the use of hedges and boosters between research articles published in Spanish and Arabic academic journals in the discipline of linguistics?
- Are these differences/similarities attributable to cultural or disciplinary influences?

Based on a corpus of 90 articles collected from 6 journals of linguistics, this study seeks to detect the similarities and differences in the use of hedges and boosters in native Spanish and native Arabic linguistics research articles. Hyland's (Hyland, 2005) taxonomy of metadiscourse markers as a model of analysis to language groups has been applied. For this purpose a list of metadiscourse categories in Spanish and Arabic has been developed. The selected texts are analyzed by means of Wordsmith Tools (5.0 and 6.0) (Scott, 2008, 2012) and then carefully checked manually in the context for metadiscourse categories. The quantitative analyses showed that the overall use of hedges and boosters in Spanish research articles is higher than in the Arabic ones. While the Spanish authors used in their writings significantly more hedges than boosters, on the contrary their Arab colleagues used more boosters than hedges. This study has showed important cross-cultural, cross-linguistic, and genre-related differences in the use of hedges and boosters. The results are especially helpful for Spanish and Arabic as a second language teaching situations. When and if differences are found to exist across texts and cultures, they can then be explained to students.

Portuguese multiword expressions: data from a learner corpus

Sandra Antunes
Centro de Linguística da Universidade de Lisboa
sandra.antunes@clul.ul.pt

Amália Mendes
Centro de Linguística da Universidade de Lisboa
amalia.mendes@clul.ul.pt

1 Introduction

The proper usage of Multiword Expressions (MWE), i.e., sequences of words with a syntactic and semantic cohesion (Mel'cuk, 1984; Sinclair, 1991, Cowie, 1998; Sag et al., 2002) is crucial in L2 studies. Indeed, L2 learners frequently struggle to choose the right combination of words and eventually produce errors related to the lexical-grammatical, semantic or stylistic aspects of MWE (Nesselhauf, 2004; Gilquin, 2007; Granger and Paquot, 2010; Paquot, 2013).

Our paper focuses on the use of MWE in a subset of COPLE2, a new learner corpus of Portuguese L2, and addresses the following issues: (i) how significant is the difficulty for the learners to produce MWE; (ii) what are the major errors students make when dealing with constrained expressions.

2 Corpus constitution

Our analysis is based on data from the written register of COPLE2[111], which is composed by: (i) 966 free handwritten essays from different genres (the most frequent being opinion), collected in evaluation tests; (ii) 424 students (18-40 years); (iii) 14 different mother tongues; (iv) all levels of proficiency (the most frequent being elementary). The corpus will be lemmatized and annotated with information on PoS and error type (Nicholls, 2003; Dagneaux et al., 2005).

We restrict our analysis to learners of Portuguese with Spanish, English and Chinese as L1 (cf. Table 1).

L1	Inf.	Age	Texts	Words	Proficiency
Chinese	129	22	323	57.377	Int. (34%)
English	65	24	142	21.610	Elem. (41%)
Spanish	52	29	139	21.200	Elem. (57%)
TOTAL	**246**	**25**	**604**	**100.195**	**------**

Table1: COPLE2 subcorpus

[111] http://www.clul.ul.pt/en/research-teams/547

3 Data analysis

Since all the essays were handwritten, and had to be digitalized and transcribed, the MWE were extracted and annotated during the transcription process. We organized the data according to the typology established by Sag et al. (2002), slightly adapted to our data, and, using a Contrastive Interlanguage Analysis approach (Granger, 1996), we identified different error types:

(i) Collocations (expressions semantically compositional but lexically and/or pragmatically constrained).
- Substitution for (quasi-)synonyms or words belonging to the same semantic field: #*maneiras de transporte* 'ways of transport' vs. *meios de transporte* 'means of transport' (Chinese).
- Substitution for phonologically or morphologically similar words: #*comida populosa* 'populous food' vs. *comida popular* 'popular food' (Chinese).
- Substitution for periphrasis or semantically related words: #*as diferenças e as coisas iguais* 'the differences and the equal things' vs. *as diferenças e as semelhanças* 'the differences and the similarities' (Chinese); #*animais preciosos* 'precious animals' vs. *animais em vias de extinção* 'endangered species' (Chinese).
- L1/L2 transfer at both lexical and syntactic levels: #*parada de metro* 'subway *parada'* vs. *estação de metro* 'subway station' (Spanish); #*especialistas biológicos* 'biological experts' vs. *especialistas em biologia* 'experts in biology' (Chinese). The last example shows that Portuguese non-predicative adjectives pose restrictions regarding the nouns they modify, requiring prepositional phrases.
- Mismatch of the copulative verbs *ser* and *estar* 'to be': #*estamos cruéis* vs. *somos cruéis* 'we are cruel' (English).
- Transposition of semantic relations: #*fechadura nórdica* 'Nordic closenness' in contrast with *abertura nórdica* 'Nordic openness' (English).

(ii) Light verbs constructions (as these verbs carry no significant meaning, the students frequently use them interchangeably): #*dar uma grande influência* 'to give a large influence' vs. *ter uma grande influência* 'to have a large influence' (Chinese).

(iii) Lexically-syntactically fixed expressions.
- Lexical mismatch: #*dia com dia* 'day with day' vs. *dia a dia* 'day after day' (English).

- L1 transfer: #*música viva* ‘live music’ vs. *música ao vivo* (English).

(iv) Routine formulae.

- L1 transfer (#*sem outras coisas para reclamar* ‘there being no other things to complaint’ vs. *sem outro assunto de momento* ‘there being no other matter to discuss’ (Chinese).

(v) Idiomatic expressions.

- Substitution for semantically related words: #*faca sempre tem dois lados* ‘knife always has two sides’ vs. *faca de dois gumes* ‘double-edged sword’ (Chinese).

4 Conclusion

Our data show that collocations are especially difficult for learners of Portuguese L2 because, even though they are semantically compositional, they pose degrees of restrictions that are not easily acquired, generating obvious errors. The few cases of idiomatic expressions in our corpus are also problematic. A possible explanation for their low frequency is that learners have elementary proficiency and are not yet familiarized with them. To target this subtype, other methods, such as translations or elicitation tests, would be required.

L1/L2 transfer plays an important role in the students’ productions and is particularly noticeable in expressions with equivalent forms in their L1. We identified cases of transfer of lexical units (either in their native language or adapted to Portuguese), syntactic constructions and register.

We believe that a clear description of the categories of MWE and the identification of factors that correlate with the learners’ difficulties may be the key to their lexical accuracy. It is our aim to provide such a typology for Portuguese.

References

Cowie, A. P. 1998. *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press.

Dagneaux, E., Denness, S., Granger, S., Meunier, F., Neff, J. and Thewissen, J. (eds.) 2005. *Error Tagging Manual. Version 1.2*. Centre for English Corpus Linguistics. Université Catholique de Louvain. Belgium.

Gilquin, G. 2007. “To err is not all. What corpus and elicitation can reveal about the use of collocations by learners”. *Zeitschrift für Anglistik und Amerikanistik*, 55.3. Pp. 273-291.

Granger, S. 1996. “From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora”. In K. Aijmer, B. Altenberg and M. Johansson (eds.) *Languages in Contrast. Text-based cross-linguistic studies*. Lund Studies in English 88. Lund: Lund University Press. Pp. 37-51.

Granger, S. and Paquot, M. 2010. “Customising a general EAP dictionary to meet learner needs”. In *eLexicography in the 21st century: New challenges, new applications*. Proceedings of ELEX2009. Cahiers du CENTAL N°7. Louvain-la-Neuve, Presses universitaires de Louvain.

Mel’cuk, I. 1984. *Dictionnaire explicatif et combinatoire du français contemporain*. Les Presses de L’Université de Montréal. Montréal. Canada.

Nesselhauf, N. 2004. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins Publishing Company.

Nicholls, D. 2003. “The Cambridge Learner Corpus - error coding and analysis for lexicography and ELT”. In Archer, D., Rayson, P., Wilson, A. and McEnery T. (eds.) *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster University (UK). University Centre for Computer Corpus Research on Language. 28-31 March. Pp. 572-581.

Paquot, M. 2013. “Lexical bundles and L1 transfer effects”. *Language Learning and technologt 14(2)*. Pp. 30-49.

Sag, I., Baldwin T., Bond F., Copestake A. and Flickinger D. 2002. “Multiword Expressions: A Pain in the Neck for NLP”, in A. Gelbukh (ed.) *Proceedings of CICLing-2002*, Mexico City, Mexico. Vol. 2276, pp. 1-15.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Multi-modal corpora and audio-visual news translation: a work in progress report

Gaia Aragrande
University of Bologna
gaia.aragrande@studio.unibo.it

1 Audio-visual news translation: a corpus based approach.

News discourse has been investigated, primarily in its written environments, by numerous scholars both in Translation Studies (e.g. Bassnett, 2005; Bassnett and Conway, 2006; Bassnett and Bielsa, 2009; Van Doorsaeler, 2012; Schäffner, 2012) and in Discourse Analysis (e.g. Morley and Bailey, 2009; Partington et al., 2013). The translational perspective of audio-visual journalism, however, has been rarely taken into account, especially from a corpus-based stance.

Audio-visual news can originate from many different sources, and the transformations inserted by news-makers often concentrate mainly on the audio-track. In foreign news reporting, therefore, translation plays a major role, which is rarely acknowledged explicitly in media studies.

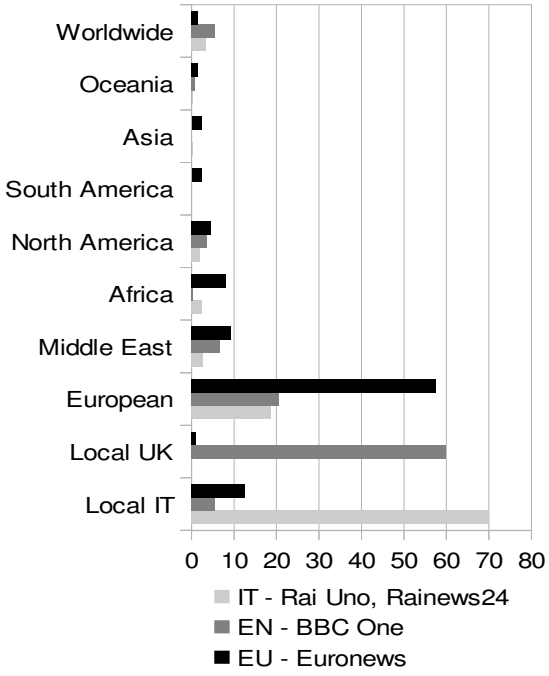


Figure 1: News Areas in Italian, English and European[112] data sets

Following the work of Tsai (2005, 2006, 2010, 2012), the aim of this study is to investigate how translation impacts on the delivery of journalistic messages in broadcast audio-visual events, through the use of a small multi-modal corpus.

The data sampled for the corpus come from three different sources. The multilingual and translational core of the corpus is represented by *Euronews* online video-news [113] in its English and Italian versions.[114]

Complementing this small parallel corpus, two comparable corpora of newscast recordings gathered from three monolingual channels were built. The channels sampled for the comparable corpora are, for the English component, *BBC One* newscasts (six and ten p.m.), and, for the Italian component, *Rai Uno* evening newscast and *Rainews24* afternoon news-updates (one and three p.m.). The recordings followed a weekly schedule and went on for two months (15/12/2015-22/02/2015), this material will eventually merge in the final corpus (see table 1 for the corpus' final composition).

The sub-corpora will be POS-tagged and annotated with speaker details and information about journalistic content (including geographical area of interest and main topic), allowing for the creation of on-the-fly thematic sub-corpora. They will also be tagged for audio-visual events that complement or interfere with the reporting activity.[115]

Sub-corpora	Comparable		Parallel	
Components	IT	EN	IT	EN
Num. of texts	120	80	390	390
Text average length (approx.)[116]	8,000	6,000	250	250

Table 1: Multi-modal corpus final composition.

2 Pilot corpus: testing phase.

The last two weeks of the recordings have been entirely transcribed and partially tagged [117] to develop a pilot corpus, which serves as methodological instrument to forestall pitfalls that might affect the construction of the final corpus, as well as to provide material for a case study, yielding initial insights about the potential value of the final corpus for discourse and translation studies.

The chart in figure 1 allows one to observe what

[112] Due to EN and IT components' symmetry in the European data set, in both charts their data were considered together.

[113] http://it.euronews.com/notizie/telegiornale/

[114] Although the corpus only includes the English and Italian versions, the channel provides rolling news in thirteen different languages, making translation one of its flagship features.

[115] E.g., off-screen narrating voices or visual support items such as slides and pictures.

[116] The average text lengths have been calculated on the word count of a two weeks transcribed sample.

[117] With reference to speakers, area and topic of the reported news and for audio and visual (or both) events that took place on the screen.

geographic areas are prioritized by different broadcasting channels in the three data sets. The English and Italian data sets prioritize domestic news coverage, both devoting around 20% of the entire coverage to EU-related matters, and both completely ignoring[118] some non-EU areas. Instead, the European data set seem to be representative, albeit in different percentages, of all the identified news areas.

In the chart in figure 2, which represents the topic composition within the European news area in the three data sets, the situation appears to be similar in the two monolingual data sets, with the Italian component covering slightly more topics than the British one.

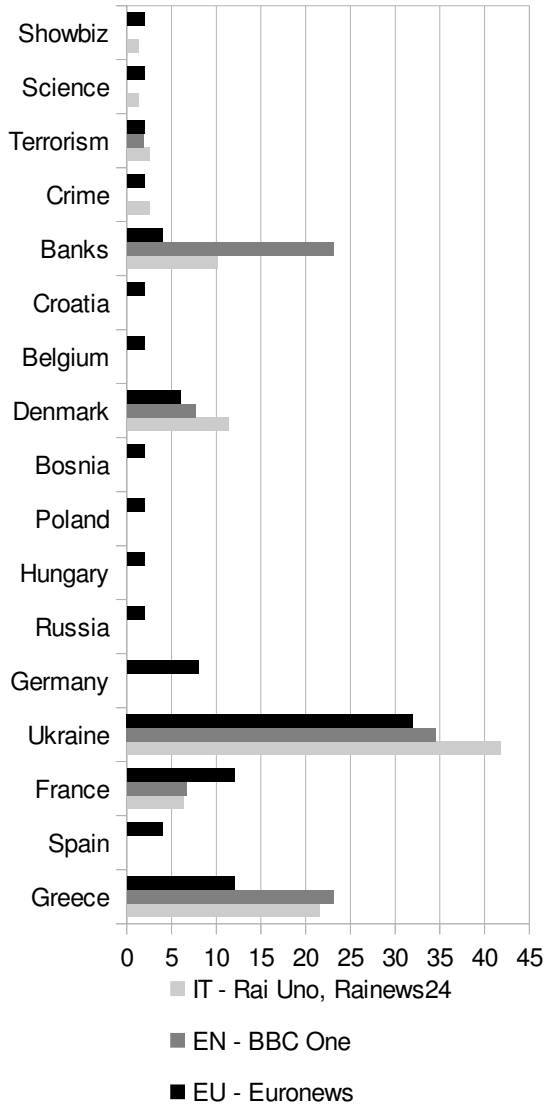


Figure 2: European News Area composition in Italian, English and European data sets.

The European data set instead seems to better represent Europe's geographic mosaic, without ignoring the urgency of newsworthy events, as shown by the partial overlapping among the data sets in representing the most prominent news items.

The whole pilot corpus contains 860 news reports, 9.65% of which are dedicated to the Ukraine's crisis. For this reason, Ukraine-related news reports have been chosen as the subject of a case study, whose preliminary results show how the search item "*ucra*/ukra**" is relatively frequent, as indicated in table 2.

	Types	Tokens	Search item f %	
			IT: "*ucra**"	EN: "*ukra**"
BBC	5,888	60,117		1.15
RAI	11,754	96.321	1.22	
Euronews EN	3,123	12,522		1.6
Euronews IT	3,717	13,007	1,45	

Table 2: Types, tokens and "*ucra*/ukra**" frequency counts[119] of the four data sets.

The analysis of Ukraine-related news focusses both on translation processes in the parallel sub-corpus, and on its coverage in the comparable sub-corpora. The audio-visual components are also taken into consideration, in order to evaluate their contribution to the news information's delivery.

References

Bartrina, F. 2004. "The Challenge of Research in Audiovisual Translation". In Orero, P., (ed.) *Topics in Audio-visual Translation*, Amsterdam/Philadelphia: John Benjamins.

Bassnett, S. 2005. "Bringing the News Back Home: Strategies of Acculturation and Foregnisation". *Language and Intercultural Communication*, 5 (2): 120-130.

Bassnett, S., Conway K. (eds) 2006. *Translation in Global News – Proceedings of the conference held at the University of Warwick, 23 June 2006*, Coventry, UK: University of Warwick, Centre for Translation and Comparative Cultural Studies.

Bassnett, S., Bielsa, E. 2009. *Translation in Global News*. New York: Routledge.

McEnery, T., Xiao, R., Tono, Y. (eds) 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.

Morley, J., Bayley, P. (eds) 2009. *Corpus-Assisted Discourse Studies on the Iraq Conflict: Wording the War*. New York: Routledge.

[118] This is striking considering that the Italian data set includes a rolling news channel (*Rainews24*), which clearly has more time of news broadcasting than *BBC One*.

[119] Software used: AntConc 3.4.3, http://www.laurenceanthony.net/software.html

Partington, A., Duguid, A., Taylor, C. 2013. *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS).* Amsterdam: John Benjamins.

Schäffner, C. 2012. "Rethinking Transediting". *Meta* LVII (4): 866-883.

Stetting, K. 1989. "Transediting – A new Term for Coping with the Grey Area Between Editing and Translating". In Hale, G. (ed.), *Proceedings form the Fourth Nordic Conference for English Studies.* Copenhagen: University of Copenhagen, 372-382.

Tsai, C. 2005. "Inside the Television Newsroom: An Insider's View of International News Translation in Taiwan". *Language and Intercultural Communication*, 5 (2): 145-153.

Tsai, C. 2006. "Translation through Interpreting: A Television Newsroom Model" in Bassnett, S., Conway, K. (eds.): 59-72.

Tsai, C. 2010. "News Translator as Reporter". In Schäffner, C., Bassnett, S. (eds.) *Political Discourse, Media and Translation.* Newcastle Upon Tyne: Cambridge Scholars. 178-197.

Tsai, C. 2012. "Television News Translation in the Era of Market-driven Journalism". *Meta* LVII (4): 1060-1080.

Van Doorslaer, L. 2012. "Translating and Constructing Images in Journalism with a Test Case on Representation in Flemish TV News". *Meta* LVII, (4): 1046-1059.

Catachrestic and non-catachrestic English loanwords in the Japanese language

Keith Barrs
Hiroshima Shudo University
Keithbarrs@hotmail.com

1 Introduction and background

A loanword in a language may or may not have a semantic near-equivalent expression made up from resources in the language's native lexicon. According to a recently-coined terminology, if a near-equivalent exists then the loanword is a non-catachrestic innovation and if not, it is a catachrestic innovation (Onysko & Winter-Froemel 2011). A catachrestic loanword works mainly to fill a lexical gap opened up by the introduction of a novel object, concept or idea from the SL. It is used for its literal, direct, denotative meaning. A non-catachrestic loanword is different in that it sits alongside one or more semantic near-equivalent terms, becoming one member of a near-synonymous pair or group. It is used primarily for its implied, suggestive, connotative meaning; an effect produced by the selection of the loanword over the near-semantic equivalent. In other words, catachrestic loanwords represent the normal, literal way of speaking about the objects and concepts concerned, while non-catachrestic loanwords "are typically used to express additional pragmatic meanings" (Onysko & Winter-Froemel 2011 p 1555).

Focusing on non-catachrestic loanwords in particular, investigating how they are used within a language can reveal the motivations of why they were selected ahead of the semantic near-equivalent terms. This can show the different kinds of pragmatic effects that they contribute to a language. Onysko and Winter-Froemel (2011) were the first to conduct such a study, using their newly-coined terminology of catachrestic and non-catachrestic innovations. They analysed one hundred highly frequent anglicisms in German, drawn from a newsmagazine corpus, and classified them as catachrestic and non-catachrestic, before describing some of the pragmatic effects generated by the categories. In the Japanese context, where tens of thousands of English words have been borrowed, the importance of understanding the semantics and pragmatics of these loanwords is similarly acknowledged, especially in the areas of lexicography and linguistics (Irwin 2011), language education (Daulton 2008; Inagawa 2010; Ringbom 2007), and cultural studies (Kay 1995; Loveday 1996; Stanlaw 2004). However, to date there has

been no large-scale study of the semantic/pragmatic behaviour of English loanwords in Japanese.

To address this research gap, a similar study to that of the anglicisms in German was planned for English loanwords in Japanese. However, an important difference is that the final phase of the research (planned for late 2015) is to use a corpus-based methodology for the analysis of semantic and pragmatic effects. This is to be done with the 'sketch-diff' tool in the Sketch Engine corpus query system, comparing the behaviour of the non-catachrestic English loanwords with their semantic near-equivalent (non English-derived) terms in Japanese. In order to do this, a pool of non-catachrestic English loanwords needs to be created. The initial plan was to follow Onysko and Winter-Froemel's model and extract the most frequent one hundred loanwords from a corpus. Then a similar method of categorisation to theirs would be applied. However, as Onysko and Winter-Froemel themselves admit, investigating the pragmatic effects of non-catachrestic loanwords in particular can be very difficult when they have multiple senses (i.e. they are polysemous) (2011 p 1563).

This is especially problematic with corpus-based studies where the software is generally unable to effectively isolate the different senses of polysemous words. The research summarised here describes the results of the one-by-one linguistic analyses of the most frequent one hundred English loanwords in Japanese. It explains the multiple stages of filtering necessary to prepare the loanwords for analysis within the Sketch Engine's 'sketch-diff' function, and discusses how this process reduced the list from one hundred loanwords to only twenty-one. The implication is that in order to create a pool of loanwords similar to that used by Onysko and Winter-Froemel, over five hundred loanwords will have to be initially analysed, greatly altering the workload and time frame of the proposed research. As such, these results reveal important considerations that need to be recognised when undertaking corpus-based research into loanwords.

2 Methodology

The jpTenTen11 web corpus of Japanese in the Sketch Engine was chosen as the resource of loanwords because of (1) its massive size meaning it would contain a large number of loanwords and thousands of hits for each one, (2) its hybrid nature as a modern spoken/written textual medium, and (3) its ease of processing with corpus tools within the Sketch Engine. Loanwords are overwhelmingly written in katakana script, so the regular expression [ア-ん]+ was used to extract all strings of katakana from the corpus. The list was then processed using the unidic dictionary that filtered the katakana strings for those that are categorised as part of the 'foreign word' vocabulary strata. Using dictionary resources, the list was then manually filtered for only the English loanwords.

The list was then filtered for those that could be appropriately analysed by the Sketch Engine software, by isolating non-homographic loanwords from homographic ones, as individual members of homographic sets cannot be easily analysed within the Sketch Engine. The non-homographic words were then categorised as monosemous or polysemous, with the polysemous loanwords being excluded from further analysis because of similar issues to those with homographs. The non-homographic, monosemous loanwords were then investigated individually for the presence of semantic near-equivalents. For this, data from (1) a monolingual dictionary (大辞林 *daijirin*), and (2) a specialised loanword dictionary (カタカナ語辞典 *katakanagojiten*), were cross-referenced with data from (3) the thesaurus function of the Sketch Engine. This function generates a list of collocationally and grammatically similar words, statistically ranked by their similarity to the search term in shared contexts of occurrence (Kilgarriff et al., 2014). Words on this list that also appeared in the dictionary entries of the search term were judged to be semantic near-equivalents, and would make that loanword non-catachrestic. If several semantic near-equivalents had been identified in the previous stage, it was necessary to isolate a single item. This was because the final stage of the research will involve a two-word comparison of one English non-catachrestic loanword lemma and one semantic near-equivalent lemma, using the 'sketch-diff' function of the Sketch Engine.

3 Results and implications

The principal finding of the research was that if corpus tools were to be used to investigate the catachrestic/non-catachrestic categories of English loanwords in Japan, then in order to conduct a similar study to that of Onysko and Winter-Froemel (2011) it is likely that over five hundred loanwords would have to be individually analysed. This is because the majority of loanwords had to be discarded at the homography/polysemy filtering stage. In detail, 79 words out of 100 had to be discarded. Of the remaining 21, 11 were judged as catachrestic and 10 as non-catachrestic. It can be assumed that a similar pattern would be found with words beyond the most frequent one hundred, thereby greatly increasing the initial amount of loanwords needed in order to produce a sizable pool that can be (1) effectively classified as catachrestic

and non-catachrestic and (2) used within the sketch-diff function to analyse their behaviour. An additional finding related to this issue was that many distinct English words have become homographs in Japanese due to differences in the phonologies of the two languages (e.g. *bath* and *bus* are both represented as バス, *basu*, in katakana). This creates problems for the corpus analysis because it is highly complicated to isolate the different meanings represented by one word, causing issues for the interpretation of the data.

These findings suggest that in order to build a sufficient database of loanwords for a large-scale corpus-based study, it is necessary to start with a number far beyond the target number for the analysis. This is because of current limitations with corpus software in the analysis of homographic and polysemous words. This can greatly alter the workload and timescale for the research, and is therefore an important consideration to be aware of when conducting corpus-based studies of loanwords.

References

Daulton, F. E. 2008. *Japan's Built-in Lexicon of English-based Loanwords*. Clevedon: Multilingual Matters Ltd.

Inagawa, M. 2010. *A Corpus-Driven Study of Loanwords : Synchronic and Diachronic Change of English-Derived Words in Contemporary Japanese*. Unpublished PhD thesis, University of Queensland.

Irwin, M. 2011. *Loanwords in Japanese*. Philadelphia: John Benjamins Publishing Company.

Kay, G. 1995. English loanwords in Japanese. *World Englishes*, *14*(l), 67–76.

Kilgarriff, A., Baisa, V., Busta, J., Jakubicek, M., Kovar, V., Michelfeit, J., Rychly, P. and Suchomel, V. 2014. The Sketch Engine: Ten years on. *Lexicography ASIALEX*, *1*, 7–36. Available online at http://link.springer.com/article/10.1007%2Fs40607-014-0009-9

Loveday, L. 1996. *Language Contact in Japan: A Socio-Linguistic History*. Oxford: Oxford University Press.

Onysko, A., & Winter-Froemel, E. 2011. Necessary loans – luxury loans? Exploring the pragmatic dimension of borrowing. *Journal of Pragmatics*, *43*(6), 1550–1567.

Ringbom, H. 2007. *Cross-linguistic similarity in foreign language learning*. Clevedon: Multilingual Matters Ltd.

Stanlaw, J. 2004. *Japanese English: Language and culture contact*. Hong Kong: Hong Kong University Press.

Objective-driven development of the first general language corpus of Tamazight

Nadia Belkacem
Barcelona University
nbelkabe7@ub.edu

1 Introduction

In this paper, we discuss the structure, functionality and challenges in the development of the first ever general language corpus of Tamazight freely accessible on the internet[120], using an objective-driven approach. Tamazight is currently classified as a separate branch in the afroasiatic group of languages (Greenberg, 1955). It is the oldest language proven to exist in North Africa, still spoken today as a first language by more than 30 million people. It also has interesting peculiarities from the point of view of linguistics (Mammeri, 1974).

Our motivation for embarking on this work stems from the fact that this language did not have a general corpus accessible on the internet and from our conviction of the paramount importance of the role of corpus linguistics in the development of the lexicography of this language. Therefore, the corpus was designed primarily with the objective of satisfying the need of lexicographic applications, such as dictionary construction.

The challenges encountered in developing such a corpus are mainly due to the language being less-resourced and the relatively unstable spelling of the same words in the different texts. We provide some solutions to this problem and manage to put together a functional corpus that meets its objective for lexicographic applications such as the construction of modern dictionaries of Tamazight. We describe some of these applications and anticipate that our approach can be successfully adopted for similarly less-resourced languages.

2 Development method and challenges

The development of the corpus has taken into account the latest guidelines in corpus compilation (Atkins et al. (1992) (McEnery & Hardie, 2012), together with the latest developments in software and user interfaces (Hardie, 2012). Always constrained by the lexicographic application objective, we first proceeded to specify the different modules of our corpus as follows:

[120] http://ugriw.net

- Genres of texts that would be included in the corpus compilation, considering availability in digital format and their relevance to lexicography.
- Data structure of the corpus entries in a relational database.
- Searching and computational linguistics functions that should be provided for lexicography.
- Layout of the user interface in web-based access.
- Functionality of the corpus management interface that would streamline the process of construction and maintenance of the corpus.

Thereafter, texts in digital format were collected and introduced in the database without any embedded annotations (Garside et al., 1997). The user interface was then developed and tested against the corpus database.

We found two types of challenges. The first one has to do with the unavailability of computational linguistics resources for this language. For example, we could not annotate the texts automatically with PoS tags due to the inexistence of PoS taggers for this language. The second type of challenge was due to the instability of word spelling in a large number of texts. The impact of these problems was significantly reduced through providing the appropriate functions in the user and management interfaces.

3 Structure

The structure of the corpus in the database can be summarized as follows:

- Each text is stored as an entry in UTF-8 format with its associated metadata, which includes the genre of the text and comprehensive information about the text
- The genres of the text are listed in Table 1
- The remaining metadata is listed in Table 2

4 Functionality

The functions available to the user can be divided into two groups: lexicometry and concordances, both very useful in lexicographical work.

The **global lexicometry** function allows us to display lexical statistics about the entire corpus or part of it as delimited by the conditions on:

- the genre of the text (it's possible to select one or several genres)
- the gender of the author (man or woman, or anyone)
- the period of production of the texts (between specified years)
- the minimum size of a word

Text Genre (fiction)	Text Genre (non-fiction)
Novel	Newspaper article
Short story	Biography
Tale	Academic
Play (theatre)	Textbook
Poem	Magazine
Lyrics	Speech
Proverb	Interview
	Blog
	Internet page

Table 1: Text Genres

Metadata	Comment
Name of author	
Date of birth	Year if date unknown
Gender of author	Man/Woman/Unknown
Original author	If text is translation
Editor	
Publisher	
Source	Written/Spoken
Region	Linguistic area
Title	Text title
Original title	if text is translation
Original language	if text is translation
Spelling revision	spelling has been revised
Total words	in this text entry
Production date	
Production place	
Date published	
Place where published	
Notes	about the text

Table 2: Metadata

The results of a global lexicometry search will display a summary of the contents of the corpus in terms of the number of words for each genre of text, together with the highest frequencies. For each genre of text, the following will be displayed: number of texts

- number of words
- percentage of the number of words with respect to the total number of words in the whole corpus
- the most frequent word, together with its frequency
- relative index of the frequency, expressed in a percentage value; that is to say, what proportion the word represents with respect to the total words in that type of text
- number of distinct words
- relative index of the number of distinct words expressed in a percentage value, i.e.

its proportion relative to the total number of words

With the **frequencies** function, we can obtain a list of all distinct words in the corpus with their frequencies in decreasing order. We can restrict our search to a subset of the texts in the corpus by specifying the appropriate conditions as for the previous function. Additionally, it is possible to request the display to start from a particular frequency, e.g. frequency 1 to get all the hapax words exclusively.

With the word **frequency function**, we can directly display the frequency of a particular word or expression. Optionally, we can compare the frequencies of two words or expressions.

With the **KWIC concordances** function, we can study the way words are used, by observing the immediate textual context in which they appear (Sinclair, 1991). The most widely used format to present concordances is called KWIC, "Keyword In Context". In this format, each line displayed shows the specified word aligned in the center together with its immediate context to the left (before the word) and to the right (after the word). In that way, we can easily see the pattern of usage and identify collocations. In addition to the possibility of limiting the search to a subset of the corpus as for previously described functions, we need to specify the number of words of the left and right contexts and word searching conditions as follows:

- *exact match* of the word or expression
- word *beginning* with the specified string (for example, to study words with the same prefix)
- word which *ends* with the specified string (for example, to study words with the same suffix)
- word which *contains* the specified string (for example, to study words with the same lexical roots or the inflected forms of a word)

With the **sentences** function, we can display entire sentences which use a specified word or expression. With respect to the previous function, there is one additional parameter to specify,"Max. words", i.e., the maximum size of the sentences in terms of the number of words. This function is being used by lexicographers to extract example sentences for dictionaries.

5 User and management interfaces

The user interface accessible through the internet, implements the search functions of the corpus that we previously described. The user interaction with the corpus is based on simple forms as input, with each field having a pre-assigned default value. The output results of the search are displayed in table format and can be redirected to a file.

As for the management interface, it is also accessible through the internet and provides the following forms-based functionality:

- Input and modification of text entries
- Automated correction of spelling in texts
- Statistics of corpus usage for improvement purposes

6 Results and conclusion

The first general language corpus of Tamazight has been developed efficiently and successfully with the objective of serving the lexicography community of this language. Since its availability online, we have noticed increasing daily usage for lexicography purposes. In our opinion, we have opened a new stage in the development of lexicography of the Amazigh language by enabling the construction of its corpus-based dictionaries. We hope to continue improving this corpus by adding new texts and developing an automatic PoS tagger for texts in this language.

References

Atkins et al. (1992) *Corpus design criteria*. In Literary & Linguist Computing, 7 (1). Oxford: Oxford University Press

Garside, R., Leech, G. N., and McEnery, T. (1997) *Corpus annotation: linguistic information from computer text corpora*. London: Longman

Greenberg, J. (1955) *Studies in African Linguistic Classification.* New Haven

Hardie, A. (2012) *CQPweb - combining power, flexibility and usability in a corpus analysis tool*. International Journal of Corpus Linguistics 17 (3): 380–409.

Mammeri, M. (1974) *Tajerrumt n Tmazight* (Grammar of Tamazight). Alger: Bouchene.

McEnery, T. and Hardie, A. (2012) *Corpus Linguistics: Method, Theory and Practice.* Cambridge Textbooks in Linguistics. Cambridge:Cambridge University Press.

Sinclair, J. M. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Prescriptive-descriptive disjuncture: Rhetorical organisation of research abstracts in information science

John Blake
Japan Advanced Institute
of Science and Technology
johnb@jaist.ac.jp

1 Introduction

Writing for publication in English is an onerous task for novice researchers. A key difficulty in drafting scientific abstracts is the necessity to adhere to the generic integrity of the discourse community. This difficulty is exacerbated for writers of English as an additional language who need not only to master the relevant lexicogrammar, but also acquire the academic literacy of their particular scientific community.

Scientific research abstracts are particularly noted for their high information density (Holtz, 2009). The high frequency of nominalization (Biber & Gray, 2013) and the markedly long grammatical subjects contribute to their high lexical density (Halliday and Martin, 1993).

2 Prescriptive advice

Ethnographic advice was extracted from guidelines for journals in information science, English-language publications housed in the resource center of a scientific research institute, and the top ten hits generated in Google using various search terms. General advice on writing scientific research abstracts focusses on ensuring that abstracts show the originality, substance and importance of the research. Specific advice is often provided on the prescribed rhetorical structure of abstracts. Many sources advocated writing abstracts using the linear four-step Introduction, Method, Results and Discussion (IMRaD) model. A number of sources also recommended including the aim, purpose or goal of the research.

3 Aim

The aim of this study is to identify whether the advice proffered reflects the reality of the abstracts published in top-tier journals in the field of information science in terms of composition and sequence of rhetorical moves.

4 Method

A corpus of 500 scientific research abstracts drawn from five IEEE journals in different subdomains of information science was created. All the journals were highly rated with a mean 5-year impact factor of 3.8 and were considered by specialist informants as top-tier journals. The first 100 research abstracts published in 2012 were selected from each journal. This small balanced corpus consists of 84,652 tokens and 3047 sentences.

Each sentence was tagged with one or more rhetorical moves, namely Introduction (I), Purpose (P), Method (M), Results (R) or Discussion (D) (Hyland, 2004, p.67). Move boundaries were identified when the tags of adjoining sentences differed. The tagging was completed manually using UAM Corpus Tool version 3.0. Five specialist informants were consulted to verify the accuracy of the annotation.

5 Results

The prescriptive advice in the vast majority of published sources advocated writing abstracts in a linear format with clear demarcations between moves (e.g. IPMRD). However, the descriptive corpus results show numerous permutations of move sequences. Three features discovered in the corpus are recursivity, fronting and omission of moves.

A surprising result was the extent to which recursivity was harnessed with frequent cycling through moves (e.g. IMRMRD) which was not mentioned in any of the guidelines.

Another unexpected result was the reversal of the expected order of pairs of moves. Conventionally, method precedes results (e.g. IMRD), yet in the corpus the method move was, at times, preceded by results (e.g. IRM).

The fronting of the Discussion or Result was often combined with the omission of Introduction (e.g. RM), creating abstracts that are far removed in terms of rhetorical organisation from the prescriptive IMRD advocated in the guidelines.

6 Discussion

The plethora of permutations of moves in the corpus is in stark contrast to the prescriptive IPMRD or IMRD in the guidelines. Generally, the IMRaD model may serve as a useful pedagogic tool to help scaffold draft abstracts, but it should be considered as the starting point rather than the goal.

The cyclic nature of some experimental abstracts can be explained by the complexity of the research, and the need to provide information in a reader-friendly format. The fronting of important information enables readers to discover the key information early in the abstract. Both recursivity and fronting, however, deserve more emphasis in guidelines.

Specialist informants commented that for known unsolved problems, Introduction and Discussion moves are unnecessary and more typical of graduate student work.

A data-driven learning approach using corpus tools to investigate abstracts in their own domain could be used to raise the awareness of the variety of rhetorical structures. This hands-on approach could help novice writers more rapidly join the discourse community and realise that prescriptive rules are, at times, open to interpretation.

References

Biber, D. & Gray, B. (2013). Nominalizing the verb phrase in scientific writing. In B. Aarts, J. Close, G. Leech & S. Wallis (Eds). *The verb phrase in English: Investigating recent language change with corpora, (pp.99-132).* Cambridge: Cambridge University.

Halliday, M., & Martin, J.R. (1993). *Writing Science: Literacy and Discursive Power*. London: Falmer Press.

Holtz, M. (2009). Nominalization in scientific discourse: a corpus-based study of abstracts and research articles. In Mahlberg, Michaela and González-Díaz, Victorina and Smith, Catherine (Eds.) *Proceedings of the 5th Corpus Linguistics Conference.* Liverpool, UK.

Hyland, K. (2004). *Disciplinary Discourses: Social Interactions in Academic Writing.* Ann Arbor: University of Michigan Press.

Building COHAT: Corpus of High-School Academic Texts

Róbert Bohát
The International School of Prague
rbohat@isp.cz

Nina Horáková
The International School of Prague
nhorakova@isp.cz

Beata Rödlingová
The International School of Prague
brodlingova@isp.cz

1 Introduction

How can language and academic writing instruction be moved from the mechanistic, subjective, and teacher-centered realm to the realm of more objective, data-driven and student-centered discovery learning with room for critical thinking and metacognition? Corpus linguistics may provide the solution, as several corpus-based projects in EAL / EAP (English as an Additional Language / English for Academic Purposes) learning at the International School of Prague (ISP) seem to confirm.

2 Background

Since 2013 EAL students at ISP have engaged in a heuristic approach to academic English learning within the framework of the Applied Linguistics Project (ALP), where they studied the interaction between their mother tongue and academic language, using either basic quantitative linguistic tools for analysis, or corpus linguistics. Students who worked with corpora (InterCorp and BNC) have shown a profound understanding of collocations, relative frequencies and polysemy, as well as a number of important insights into academic English. When presenting their semantic or grammatical discoveries to their classmates, these young researchers became *de facto* co-teachers in the classroom, which resulted in a lively atmosphere of genuine academic discussion among all the students.

The fact that the discussions were based on factual evidence derived from corpora represents an innovative approach to an academic discipline that does not typically work with large sets of data and is often considered to rely largely on subjective evaluation of texts.

Corpus of high-school academic texts: rationale and design

The positive results of the corpus parts of ALP inspired the creation of a specialized Corpus of High-School Academic Texts (COHAT), providing

a high-school level addition to the repertoire of general and specialized corpora. The rationale for building a corpus of this type is twofold. First, the existing range of corpora seems to cater predominantly to higher levels of academic discourse, offering collections of university essays or academic papers. Second, most of the high-school learner corpora we found focused on identifying problem areas in non-native speaker texts; the goal of COHAT is to provide high school students with a set of successful academic English texts written by their peers that would focus on detecting patterns of correct word choice, syntax and style in students' writing. In other words, "only texts that have met departmental requirements for the given level of study" were and will continue to be included, just as in the university level British Academic Written English (BAWE) Corpus. (Alsop and Nesi 2009) This basic structure could later be enhanced by teachers' texts (e.g. samples, model essays and reports), exemplifying the ideal and allowing for a quantifiable comparison of teacher expectations with the reality of student writing.

The final product therefore aims to be subdivided into two sections, Teacher Writing and Student Writing, each of these organized by discipline, genre, grade level and the author's mother tongue (English native or non-native). This could later be expanded into a wider International School Academic English Corpus (ISAEC) with samples added from international (or other English-medium) high schools worldwide.

3 COHAT: current status

At this first stage, COHAT is small and basic: containing 50,000+ words without annotation, allowing for the following:

- concordancing
- word lists
- frequency studies, complemented with immediate context 'visible' in the concordance
- basic collocation studies.

All of the above could be used to study the typical syntax, grammar, word choice, etc. of sample texts for each subject and genre. Currently, four discipline-related categories are represented:

- English and Literature
- Social Studies
- Maths and Natural Sciences
- Creative Writing (including Speeches & Journalism).

The plan is to create a balanced corpus with similar sized collections of student and teacher writing for each high school grade level (i.e. Grade 9, 10, 11, and 12). Following this, grammatical and semantic taggers will be used to annotate/tag the anonymized texts to allow for further linguistic analysis of the data.

4 Conclusion

Within the current range of corpora available, a corpus of high-school student academic writing that focuses on exemplary student writing rather than error detection seems to be missing or not readily available. In other words, the creators of COHAT have "identified a clear gap in the research with the potential for some really interesting and useful work." (Gupta 2014)

COHAT will allow for a quantifiable and data-driven analysis and evidence-based learning in a discipline that is traditionally considered to be subjective and difficult to grasp using scientific methods. Furthermore, it will enable students to conduct their own research into the language of academic writing, highlighting the idea of heuristic and constructivist learning.

References

Alsop, S. and Nesi, H. 2009. "Issues in the development of the British Academic Written English (BAWE) corpus." *Corpora.* 4 (1): 71-83.

Gupta, K. 2014. "Corpus Linguistics MOOC: Discussion question for week 4." *Future Learn.* Lancaster University. Available online at https://www.futurelearn.com/courses/corpus-linguistics-2014-q3/steps/14848/progress?page=5#comment_2311204

Crowdsourcing a multi-lingual speech corpus: recording, transcription and annotation of the CrowdIS corpora

Andrew Caines
University of Cambridge
apc38@cam.ac.uk

Christian Bentz
University of Cambridge
cb696@cam.ac.uk

Calbert Graham
University of Cambridge
crg29@cam.ac.uk

Paula Buttery
University of Cambridge
pjb48@cam.ac.uk

1 Overview

We present the 'CrowdIS' corpora – CrowdISeng and CrowdISengdeu – being collected from English native speaker ('eng') and German/English bilinguals ('engdeu') via crowdsourcing platforms (hence 'Crowd') for a special session on 'Advanced crowdsourcing for speech and beyond' at this year's INTERSPEECH conference (IS). Efforts to collect the corpora are ongoing, and so we describe the collection methodology, our objectives for the corpora, and explain how to stay informed of developments. The corpora will be made freely available to other researchers.

2 Crowdsourcing corpora

It is well-known that building speech corpora is a time-consuming and expensive process: one estimate puts the cost of transcription at €1 per word, before the cost of any extra annotation (Ballier & Martin 2013). Presumably the main expense in this figure is researcher time – skilled labourers with accompanying overheads. We present a method of collecting speech corpora via crowdsourcing facilities, showing that we can reduce costs considerably by distributing the work among multiple online workers.

This user-generated approach to corpus building was pioneered in the first British National Corpus collection project of the 1990s, at the time with tape recorders and a £25 gift voucher [121] as payment (Crowdy 1993). The approach has been taken forward to the current 'BNC 2014' collection project, with contributors now submitting recordings made on mobile/tablet devices along with an invoice for payment [122]. A similar device-based collection procedure was used by Kolly & Leemann (in press) to gather recordings of Swiss German dialects.

To collect recordings, we used the Crowdee application designed for Android operating systems (http://www.crowdee.de), and for transcription and annotation we uploaded those recordings to CrowdFlower (http://www.crowdflower.com). Each step is further described in the next section.

3 Recordings

Our primary motivation in proposing this project was to obtain a benchmark corpus of English native speakers undertaking tasks similar to those typically contained in learner corpora; and in our case relating to certificates of business English. Hence, a majority (65%) of Crowdee funding was allocated to the recordings needed for CrowdISeng, enabling a maximum of 130 individuals to make contributions.

Participants are required to be resident in the United Kingdom, United States or Canada, and it is a stated requirement of the task that English should be their mother tongue. They are presented with various business-related scenarios (e.g. starting a business, hosting visitors, sports sponsorship), and posed five questions (or 'prompts') about each scenario.

In total, the jobs feature twenty prompts and participants are expected to produce approximately 300 seconds (5mins) of speech. Payment of €5 is awarded to workers who provide recordings of sufficient duration and quality, and who apparently meet the native speaker requirement.

The German/English task designed for the bilingual corpus (CrowdISengdeu) is the same in design as for CrowdISeng, except that participants have to be resident in Germany and need to define themselves as bilinguals with either language as mother tongue. They answer prompts about the same two scenarios in both English and German, and are expected to provide 150 seconds of English and 150 seconds of German. Funds currently allow for a maximum of 50 contributors.

4 Transcription and annotation

Approved Crowdee soundfiles are then uploaded to CrowdFlower, where workers are asked to complete four tasks:

1. confirm that there is spoken content in the soundfile;
2. transcribe the speech content as faithfully as possible, using full stops to divide the text 'so that it makes most sense';
3. write a corrected version of the transcribed text;
4. how likely they think it is that English/German is the speaker's mother tongue (scale of 1 to 5).

Each recording is transcribed and rated by two

[121] BNC spoken permissions request letter: http://www.natcorp.ox.ac.uk/corpus/permletters.html#spoken1
[122] BNC 2014: http://languageresearch.cambridge.org/index.php/spoken-british-national-corpus/5-starter-steps-to-taking-part

different crowdworkers, after which various annotation layers can be added. These include error corrections, sentence boundaries, phone alignments, part-of-speech tags and grammatical relations. All sound and text files will be made available when the corpora are released.

5 Participation and updates

At the time of writing the corpora are 30% complete. Participation and assistance with publicity is very much welcomed[123], whilst researchers interested in eventually obtaining the corpora may bookmark our reserved Speech and Language Data Repository URLs[124].

Acknowledgements

This work has been funded by Cambridge English Language Assessment, Crowdflower and Crowdee. We thank Tim Polzehl of Technische Universität Berlin for his help in designing and publishing the Crowdee jobs. We thank Wil Stevens of Crowd Flower for his assistance with the transcription/annotation jobs.

References

Ballier, N. & P. Martin (2013). Developing corpus interoperability for phonetic investigation of learner corpora. In: Díaz-Negrillo, A., N. Ballier, P. Thompson (eds.), *Automatic treatment and analysis of learner corpus data.* Amsterdam: John Benjamins.

Crowdy, S. (1993). Spoken corpus design. *Literary and Linguistic Computing* 8: 259-265.

Kolly, M-J. & A. Leemann (in press). Dialäkt Äpp: Communicating dialectology to the public — crowdsourcing dialects from the public. In: Leemann, A., M.-J. Kolly, V. Dellwo, S. Schmid (eds.), *Trends in Phonetics and Phonology. Studies from German-speaking Europe*. Bern / New York: Peter Lang.

[123] Further information at: http://apc38.user.srcf.net/outreach/#crowd

[124] http://sldr.org/ortolang-000913, http://sldr.org/ortolang-000914

Fit for lexicography? Extracting Italian Word Combinations from traditional and web corpora

Sara Castagnoli University of Bologna s.castagnoli@unibo.it

Francesca Masini University of Bologna francesca.masini@unibo.it

Malvina Nissim University of Groningen m.nissim@rug.nl

1 Introduction

It is widely acknowledged that lexicographers' introspection alone cannot provide comprehensive information about word meaning and usage, and that investigation of language in use is fundamental for any reliable lexicographic work (e.g. Atkins and Rundell 2008:53). This is even more true for dictionaries that record the combinatorial behaviour of words (Hanks 2012, Ramisch 2015:5), where the lexicographic task is to detect the typical combinations a word participates in. The issue of data sources and data selection is therefore all the more central for usage-based accounts of combinatorial phenomena.

2 Web corpora and lexicography

Web corpora have nowadays made their way in the lexicographic world. And reasonably so, were it only for the fact that traditional, pre-web, well-balanced general language corpora like the British National Corpus do not exist for all languages. In the case of Italian, for instance, no such resource is publicly available, and for several years the *La Repubblica* corpus (Baroni et al. 2004), entirely composed of articles from the homonymous daily newspaper, was the only corpus available to the scientific community. But words behave differently in different contexts of use, so a single-source corpus cannot be expected to provide all the data needed to obtain a comprehensive description of word usage (Atkins and Rundell 2008:66). In addition, even when general language, balanced resources exist, they might not necessarily be the perfect tools for lexicography, representative as they are of someone's choices and requiring constant – and costly – updates and enlargements to avoid obsolescence (Landau 2001).

These limitations may be overcome by large web-derived corpora. Even though little is generally known of their actual contents, in that they are assembled through automated procedures, these are less likely to be affected by skewing. Moreover, their size makes considerations about text selection

and worries about "bad" usage less relevant, as the impact of idiosyncrasies is expected to be negligible (Atkins and Rundell 2008:55-69).

Research done in comparing web-derived corpora with benchmark collections like the BNC is encouraging. Ferraresi et al. (2010), for instance, show that the data derived from the automatically built ukWac corpus are comparable both quantitatively and qualitatively to the data obtained from the BNC. The authors also suggest that web corpora may be more useful for a lexicographer because their larger size provides a better coverage of certain word senses and because they provide a more up-to-date snapshot of language in use.

3 Evaluating corpora for the extraction of Italian Word Combinations: a pilot study

A pre-web and a web-derived corpus for Italian – *La Repubblica* (Baroni et al. 2004) and PAISA' (Lyding et al. 2014) – are compared with respect to the task of extracting word combinations for inclusion in a combinatory dictionary.

To this purpose, combinatory information for selected Italian target lemmas (TLs) is obtained by extracting from the two corpora all their occurrences in a set of pre-defined POS sequences deemed representative of Italian Word Combinations, using the EXTra tool (Passaro and Lenci, forthcoming). Candidates are then evaluated by comparison with word combinations recorded in the entries for the same TLs in the largest existing Italian combinatory dictionary (DiCI, Lo Cascio 2013), which is essentially manually compiled.

Besides assessing which corpus shows the highest recall, manual inspection of the top candidates in both datasets is used to assess the proportion of valid word combinations that are extracted from the corpus but unattested in DiCI. This is expected to provide indications as to which corpus would ensure better dictionary coverage.

Acknowledgements

This research is carried out within the CombiNet project (*Word Combinations in Italian*), funded by the Italian Ministry of Education, University and Research.[125]

References

Atkins, B.T.S and Rundell, M. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A. and Aston, G. 2004. "Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-compliant Corpus of Newspaper Italian". In *Proceedings of LREC 2004*. 1771–1774.

Ferraresi, A., Bernardini, S., Picci, G. and Baroni, M. 2010. "Web corpora for bilingual lexicography. A pilot study of English/French collocation extraction and translation". In R. Xiao (ed.) *Using Corpora in Contrastive and Translation Studies*. Newcastle: Cambridge Scholars. 337–359.

Hanks, P. 2012. "Corpus Evidence and Electronic Lexicography". In S. Granger and M. Paquot (eds.) *Electronic Lexicography*. Oxford University Press. 57-82.

Landau, S. (2001) *Dictionaries: The Art and Craft of Lexicography*. Cambridge: CUP.

Lo Cascio, V. (ed.) 2013. *Dizionario combinatorio italiano*. Amsterdam/Philadelphia: John Benjamins.

Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell'Orletta, F., Dittmann, H., Lenci, A. and Pirrelli, V. 2014. "The PAISA Corpus of Italian Web Texts". In F. Bildhauer and R. Schäfer (eds.) *Proceedings of the 9th Web as Corpus Workshop (WaC-9) @ EACL 2014*. 36–43.

Passaro, L.C. and Lenci, A. forthcoming. "Extracting Terms with EXTra". To be presented at *EUROPHRAS 2015 - Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*. Malaga, Spain, 29 June - 1 July 2015.

Ramisch, C. 2015. *Multiword Expressions Acquisition. A Generic and Open Framework*. Berlin: Springer.

[125] http://combinet.humnet.unipi.it/.

Aspects of code-switching in web-mediated contexts: the *ELF WebIn* Corpus

Laura Centonze
Università del Salento
laura.centonze@unisalento.it

1 Introduction

Previous research in the field of web-mediated types of discourse has focused on the analysis of discourse-specific features of language (Pérez-Sabater 2012; Lee 2002), on the contrastive study of different forms of web communication (Lin and Qiu 2013), on the proximity factor in interned-based communication (Grabher and Ibert 2006) as well as on the emergence of new theoretical models of interpersonal communication and information-processing (Walther 1996; Maldonado *et al.* 2001). Notwithstanding this, although the worldwide phenomenon of social-networking involves the interaction of even larger numbers of people from all over the world, rarely does literature analyse interactions among people from different 'lingua-cultural backgrounds' (Cogo *et al.* 2011) using English as a Lingua Franca (ELF; cf. Seidlhofer 2001; Mauranen 2007; Jenkins 2007) for mutual understanding over the web.
Thanks to a combination of quantitative and qualitative analysis methods, the present research project aims at shedding light on the characteristics of endonormative varieties of ELF in asymmetric communicative contexts that are emerging as a consequence of the spreading of the social-network phenomenon.

2 The *English as A Lingua Franca in Web Interaction* (ELF WebIn) Corpus project

The *English as a Lingua Franca in Web Interaction* corpus (henceforth *ELF WebIn* corpus) is an under-construction collection of social-network interactions among individuals speaking different L1s and resorting to ELF as a means to communicate and seek advice on a variety of topics, ranging from work permits, language certificates, job vacancies to visa consulting and application forms. A breakdown of *The ELF WebIn Corpus* is provided in table 1:

3 Research methodology

For the purposes of our analysis, we identified and isolated the different types of code-switching (CS) in the corpus and had a look at their occurrences in context, by means of AntConc 3.2.4w (Anthony 2014). Before doing this, however, we had to convert files into the format which was most convenient for such an analysis: we generated a table for each corpus, which displayed the names of the group members on the left, and the content of posts on the right. Where more than one comment belonged to the same 'main post', we made a distinction between Answer 1 and Answer 2 and so on (A1, A2).

FB webpage	No. of words	Years	Topic
1st for Immigration-UK Visa Experts	852	2013-	Visa; job; residence
Global Visa Support	2,860	2013-	Visa; job; residence; education
UK Visa and Work Permit	4,627	2013-	Visa; job
USA Visa Experience	7,849	2013-	Visa; job; education; info
USA Visa Experiences, Questions and Confessions	2,004	2013-	Visa; job; confessions
Tot.	17,192		

Table1: Breakdown of The ELF WebIn Corpus.

4 Main findings

It was found that the use of CS is present in some specific sections of our study corpus (namely *Global Visa Support* and *UK Visa and Work Permit*), and the context in which it occurs is predominantly intersentential. Moreover, most instances of CS are accompanied by the *po* particle occurring 21 times; its closest correspondence in English is 'please' as well as 'Sir' and belongs to Filipino, the standardized form of Tagalog, i.e. the language spoken in The Philippines, and is used as a politeness formula to show respect towards elderly people as well as to those in authority. The breakdown of the *po* particle shows it is used mainly in an intersentential position and that it tends to elicit preferred answers when asking for visa-related information.

CS	Preferred	Dispreferred
Intersentential	18 (21)	3 (21)
Final/initial	6 (6)	0
No CS	11	25

Table2: CS and the *po* particle.

5 Conclusions

In the present study it has been shown that CS is not simply used in order to compensate for a lack of linguistic competence on the part of the interactant, but – as also stated by Watts (2003) - becomes a sociolinguistic practice serving pragmatic needs, since the speaker wants to reach a wider audience by means of the same language that is shared by the Facebook community but, at the same time, the use of politeness formulae (e.g. *po*) proves to be a very successful strategy for eliciting positive answers on the counterpart. This inevitably adds a new function to the already-existing pragmatic nuances that CS may acquire in context, especially when seen in multicultural settings where meaning has to be negotiated and CS becomes a doubly valuable way in order to both bridge the gap between different linguistic competences and, at the same time, obtain the desired answer.

References

Anthony, L., 2014. *Antconc 3.2.4w*, Tokyo, Japan: Waseda University. Available online at http://www.antlab.sci.waseda.ac.jp/ .

(eds.) Cogo, A., Archibald, A. and J. Jenkins, 2011. *Latest trends in ELF research*. Cambridge: Cambridge Scholars Publishing.

Grabher, G., J. Maintz, 2006. "Learning in personal networks: collaborative knowledge production in virtual forums". *Working Papers Series*, Centre on Organizational Innovation, Columbia University: 1-12.

Jenkins, J., 2007. *English as a Lingua Franca: Attitude and Identity*. New York: Oxford University Press.

Lee, C. K. M., (2002). "Literacy practices in computer-mediated communication in Hong Kong". *The Reading Matrix* (2): 1-25.

Lin, H., L., Qiu, 2013. "Two sites, two voices: linguistic differences between Facebook status updates and Tweets". *Lecture Notes in Computer Science* (8024): 432-440.

Maldonado, G. J., Mora, M., García, S., P., Edipo, 2001. "Personality, sex and computer-mediated communication through the Internet". *Anuario de Psicología, Vol. 32 (2)*: 51-62.

Mauranen, A., 2007. "Hybrid Voices: English a the Lingua Franca of Academics". *Language and Discipline Perspectives on Academic Discourse*, K. Flottum (ed.), Newcastle, UK: Cambridge Scholars Publishing: 243-59.

Pérez-Sabater, C., 2012. *The Linguistics of Social Networking: A Study of Writing Conventions on Facebook.* Available online at http://www.linguistik-online.de/56_12/perez-sabater.html

Seidlhofer, B., 2001. "Closing a conceptual gap: the case for a description of English as a lingua franca". *International Journal of Applied Linguistics* (11): 133-158.

Walther, J. B. 1996. "Computer-mediated communication: impersonal, interpersonal, and hyperpersonal interaction". *Communication Research* (23): 3-43.

Watts, R. J., 2003. *Politeness*. Cambridge: Cambridge University Press.

Semantic relation annotation for biomedical text mining based on recursive directed graph

Bo Chen
Hubei University of Art and Science
bochen@whu.edu

Chen Lyu
Wuhan University
lvchen1989@whu.edu

Xiaohui Liang
Wuhan University
1504719992@qq.com

1 Introduction

Currently dependency structure is one of the most popular representation methods. However, many problems are encountered in parsing biomedical text, in which there are many special sentence patterns, such as postpositive attributive, inverted sentences, the complex noun phrase, the verb-complement structure, etc. It is difficult to find the correct head, which leads to errors extracting entity relations.

We put forward a new method "recursive directed graph" for parsing biomedical text. In previous work, we already built a large-scale semantic resource with 30 000 Chinese sentences with feature structure in three years. It enriches Chinese Semantics resources [9]. It is an attempt to use "recursive directed graph" in annotation of English biomedical text.

2 Annotation with recursive directed graph

Generally, a phrase or sentence may be expressed as a collection of feature structures, and a feature structure is represented as a triple:

[Entity, Feature, Value]

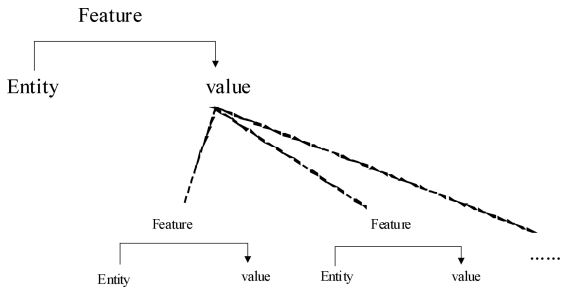


Fig. 1 Feature structure: recursive directed graph

Example 1: **Regulation** of T1 expression during induction of monocytic differentiation by okadaic acid

(1) is the title of a paper, which is a complex noun phrase with serial nouns. The sentence structure is more complex, in which the semantic relations are interrelated and complex. (1) can be described by 6 triples:

Triple1-1: [regulation, during, induction];
Triple1-2: [regulation, of, expression];
Triple1-3: [induction, of, differentiation];
Triple1-4: [differentiation, by, okadaic acid];
Triple1-5: [expression, , T1];
Triple1-6: [differentiation, , monocytic].

In triple1-2, "*expression*" is the value of the entity "*regulation*", meanwhile, in triple1-5, "*expression*" is the entity, whose value is "*T1*". And "*differentiation*" has the same situation. Therefore, in feature structure model, one node can be multiple-semantic relations node. Figure 2 is the feature structure graph of (1).

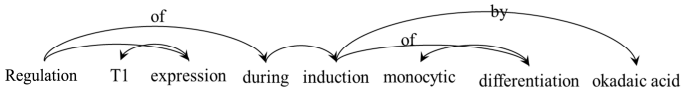


Fig. 2 The feature structure graph of (1)

3 Semantic annotation of postpositive attributive sentence patterns in biomedical text

Postpositive attributive sentence pattern in biomedical text is very common. In syntax, there are three types.

Example 2: In contrast, in a number of multiple myeloma cell lines, representing differentiated, plasma cell-like B cells, **PU.1 DNA binding activity**, **mRNA expression, and Pu box-dependent transactivation** were absent or detectable at a very low level.

In (2), it is hard to ensure the objects of the postpositive attributive verb "*binding*". It is just "*activity*", or "*activity*, *mRNA expression*", or "*activity*, *mRNA expression*, *and Pu box-dependent transactivation*". According to the semantic annotation, the subject of "*binding*" is "*DNA*", its object should be "*activity*". The postpositive attributive in (2) can be described by three triples, Figure 3 is the feature structure graph of (2).

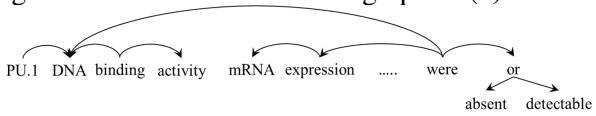


Fig. 3 The feature structure graph of (2)

Postpositive attributive is more error-prone than other sentence patterns. We just annotated 82 postpositive attributive sentences, and summarized the main three types. Using feature structure model can resolve these problems, and can represent more semantic information from biomedical texts than traditional dependent structure.

4 Conclusion

We selected 113 text materials, 11 abstracts from BioNLP'09 ST, and 102 documents from BioNLP2013 GE task. We construct a small biomedical semantic resource with 906 sentences, and focus on annotating semantic relations of sentences. We compare feature structure with Stanford parser to parse the example sentences.

The novel model "feature structure" that we put forward is formalized "recursive directed graph" for the semantic representation. It is a successful attempt to use the method in biomedical text. In future work, we will expand the biomedical corpus. Compared with other models, feature structure is more suitable for extracting biomedical complex semantic relations, and can represent more semantic relations and allows multiple links. According to the results, labeling with feature structures is much more expeditious and effective than dependency structures. In the application, our research is significant to biomedical text mining by providing rich semantic information. The resource can be used directly to relation extraction, event extraction, and automatic question and answering.

Acknowledgements

Supported by the National Natural Science Foundation of China (61202193, 61202304), the Major Projects of Chinese National Social Science Foundation (11&ZD189), and the Chinese Postdoctoral Science Foundation (2013M540593, 2014T70722).

The building of a diachronic corpus of conceptual history of Korea

Ji-Myoung Choi
Yonsei University
amancio.choi@gmail.com

Beom-Il Kang
Yonsei University
Kangbeomil@gmail.com

This paper reports on the compilation of a subcorpus of the Hallym Corpus of Korean Conceptual History (HCKCH), which aims to serve as the large-sized and linguistically analysed digital resource for diachronic studies on conceptual lexicon in the Korean history. Inspired by Koselleck's work (1975), the studies intend to examine the formation and evolution of a dozen of pivotal concepts, such as *empire, liberty*, and *labour*, in the Korean society. To do "the practice of conceptual history" (Koselleck, 2002) in the Korean and East-Asian contexts, the HCKCH corpus is being built as a specialized diachronic corpus spanning over 500 years from the late 14th century to the present.[126].

This subcorpus building has several implications for the Korean corpus research as well as the conceptual history research itself. Firstly, this corpus contains articles of all the major magazines in the critical period of the Korean history, i.e. from the late 19th century to the end of the Japanese colonial rule in 1945. This period is said to be a bridge between the feudalism of the Choseon Dynasty (from the late 14th to the late 19th century) and the modern republican society (from 1946 to the present). During this period, new types of magazines were published by a variety of newly formed political and intellectual organisations to promote their ideas and agendas (Yim, 2008).

Secondly, this corpus is the first large-sized and linguistically processed corpus of the period of 1897-1945 as far as we know: the corpus comprises of 14,606 text files, 11,133,841 eojeols[127], and 24,481,836 words. The language of this period contains a variety of orthographic variants and morphological complexity, just like the socio-political turmoil of the historical period. In addition to this orthographic complexity, many sentences have Chinese characters mixed with Korean characters within the same eojeol. It has made

[126] Funded by the government and managed by the Hallym University, this research project has recently expanded into the 'project of revealing intercommunication of basic concepts in East Asia'

[127] 'Eojeol' is one of the components in Korean orthography. An eojeol is a sequence of more than one 'umjeol' (i.e. a syllable) and is separated by spaces. An eojeol can represent more than one lexeme.

linguistic processing, i.e. segmentation and pos-tagging, of the texts extremely time-consuming and labour-intensive. To resolve these problems, a corpus processing pipeline has been developed in python and java languages. It is composed of several processing modules from the initial data collection to data standardisation to the final production of the structured corpus database. Especially, separate segmentation and tagging algorithms are developed for the Chinese character sequences and Korean character sequences, and the two processing results are combined.

For the Chinese character sequences, a lexicon-based segmentation algorithm has been developed based on the lexicon list of more than 300,000 entries. This algorithm extracts all n-grams of the Chinese characters and match them against the lexicon list. For Korean, the old Korea characters and pre-modern orthographic variants have been normalized into modern forms to improve automatic processing performance of the segmenter and tagger. Instead of using a special tool like VARD2 (Baron and Rayson 2009; Hendrickx and Marquilhas 2011), a spelling normalization module has been developed based on pre-modern character sequence rules as the latter operates on Korean faster and more flexibly. Firstly, the backbone normalization rules of about 1,650 are written by examining all the character sequences of the most frequent 10,000 eojeols occurring more than 100 times. Next, all the uni-, bi-, and trigrams of character sequences of all the eojeols are extracted from the Hallym corpus and the modern Korean corpus (*Sejong Corpus* [128]) respectively. Then their presence/absence and relative frequency are compared to each other. All in all, about 2,500 sequence normalisation rules have been written and applied to the entire corpus data.

Thirdly, this corpus is the first large-sized Korean corpus which is fully TEI-XML compliant. Previously-built Korean corpora adopted unique data formats which make them incompatible with standard ones due to the linguistic characteristics of Korean, in particular the presence of 'eojeol'.

"Table 1" and "Table 2" show the basic statistics of the resulting corpus.

unit	counts
sentence	953,541
eojeol (a sentence component)	11,131,841
morpheme (or word)	24,481,836

Table 1: The statistics of the text structure units

word class	tokens	types
(common) noun	7,266,894	226,257
(lexical) verb	2,727,350	17,663
adjective	682,155	5,986
adverb	969,242	10,227

Table 2: The statistics of major word classes

For further research, we plan to do several macro-analyses using text-mining techniques, first with topic modelling, to draw the conceptual map of this historical period. The macro-analyses could illustrate intellectual structure hidden to the researchers' naked eye, and coupled with corpus linguistics techniques, will help researchers find their ways through enormous textual data, and make it possible to compare socio-political and ideological evolutions of certain concepts synchronically and diachronically between Korea and other Asian countries, and between Korea and Western countries. In addition to the macro-analysis, the post-editing and refining of segmentation and POS tagging results will be carried out, in particular for the Chinese character sequences. The post-editing can further improve the algorithms for automatic segmentation and tagging for the pre-modern Korean language.

References

Baron, A. and Rayson, P. 2009. "Automatic standardization of texts containing spelling variation: How much training data you need?". In *Proceedings of Corpus Linguistics 2009*. University of Liverpool, Liverpool.

Hendrickx, I. and Marquilhas, R. 2011. "From Old Texts to Modern Spellings: An Experiment in Automatic Normalisation". *Journal for Language Technology and Computational Linguistics* 26(2):65–76

Koselleck, Reinhard. 2002. *The Practice of Conceptual History: Timing History, Spacing Concepts*. Translated by Todd Samuel Presner. Stanford: Stanford University Press.

Koselleck, Reinhard. 1975. "The Temporalisation of Concepts". Unpublished paper, Paris. Available at http://www.jyu.fi/yhtfil/redescriptions/Yearbook%201997/Koselleck%201997.pdf

Sangseok, Yim. 2008. *The Formation of the Korean and Chinese mixed-up style in the 20th Korean language*. Seoul: Jisik-Sanup Publications Co., Ltd.

[128] https://ithub.korean.go.kr/user/main.do
http://www.sejong.or.kr/

Top-down categorization of university websites: A case study

Erika Dalan
University of Bologna
erika.dalan@unibo.it

As student and staff mobility is moving high on the European education political agenda, it becomes imperative for universities involved in the Bologna Process and the EHEA (European Higher Education Area) to adapt their academic programs to international needs and invest in successful communication strategies, providing international students and other interested parties with easy access to complete and transparent information (Vercruysse and Proteasa 2012). Producing web-based contents in English is one of the strategies adopted to communicate with the international community, since "the use of English as a lingua franca has become accepted as a fact of life in European higher education" (Mauranen 2010). This could affect the way in which university websites in English are conceived and structured from a genre perspective, regardless of the country of origin and the variety of English adopted – native English and ELF (English as a Lingua Franca).

To gain knowledge as to how European universities structure their web-based contents in English, we conducted a genre-driven study building on the theoretical approach developed by Swales (1990) and Biber et al. (2007), with the ultimate goal of integrating genre-related data into corpora and conducting a corpus-driven analysis of institutional-academic texts. Corpus-based studies and genre-driven discourse analyses are not easily reconcilable as the former are often associated with quantitative measures describing widespread patterns of language, whereas the latter are based on qualitative and detailed analyses of a small sample of texts (Biber et al. 2007). However, applying corpus linguistics' techniques and methods to genre studies has significant benefits regarding the representativeness of the analyzed population. Furthermore, coding texts in a corpus with qualitative/interpretive data (e.g. communicative purposes) might provide new insights for a deeper understanding of linguistic differences/similarities across varieties of a language, e.g. native English and ELF.

For the above purposes, we carried out a top-down categorization of a small sample of ELF and native English websites (taken as a case study), adapting Swales' move analysis to web-based macro-genres. Move analysis, originally performed on single texts, i.e. research articles, has been conducted on the whole university website; the rationale behind this is that university websites could be associated to the concept of colony, made up of embedded component parts being themselves colonies (Hoey 2001). Similarly, university websites, taken as a whole, are a unique genre serving a set of communicative purposes, each of them realized, recursively, through single website portions, i.e. webpages and webpage sections, roughly corresponding to Swales' moves and steps. The "about us" pages, for example, aim at presenting universities to their stakeholders through a number of strategies/steps, e.g. describing university history, giving information on governing bodies and administrative structures, providing contacts. Not only does description of university websites contribute to research on genre by providing a new and dynamic concept of web-based genre analysis, it crucially enhances corpus linguistics' techniques by (manually) coding texts with genre-related metadata. After assigning genres to the webpages and using this information to construct subcorpora, the latter are analyzed to identify typical micro-features as well as similarities and differences between ELF and native English university (sub-)genres.

This study is part of a wider Ph.D. project, which in the next future aims to combine top-down and bottom-up procedures for classifying university webpages. Building on results from the move analysis and top-down categorization, we will carry out experiments for automating the process by bootstrapping our manually coded corpus and using internal criteria to conduct a bottom-up automatic categorization along the lines of Forsyth and Sharoff (2014), eventually combining both perspectives.

References

Biber, D., Connor, U. and Upton, T. 2007. *Discourse on the move: using corpus analysis to describe discourse structure*. Amsterdam: John Benjamins.

Forsyth, R.S. and Sharoff, S. 2014. "Document dissimilarity within and across languages: a benchmarking study". *Literary and Linguistic Computing*, 29 (1): 6-22.

Hoey, M. 2001. *Textual interaction: an introduction to written discourse analysis*. London: Routledge.

Mauranen, A. 2010. "Features of English as a lingua franca in academia". *Helsinki English Studies* 6: 6-28. HES Special issue on English as a Lingua Franca.

Swales, J. M. 1990. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Vercruysse, N. and Proteasa, V. 2012. *Transparency Tools across the European Higher Education Area*. The Flemish Ministry for Education and Training.

Mind-modelling literary characters: annotating and exploring quotes and suspensions

Johan de Joode
University of Nottingham
johan.dejoode@nottingham.ac.uk

Michaela Mahlberg
University of Nottingham
michaela.mahlberg@nottingham.ac.uk

Peter Stockwell
University of Nottingham
peter.stockwell@nottingham.ac.uk

1 Background

The CLiC Dickens project studies how readers mind-model fictional characters (Stockwell 2002). When textual cues trigger and interact with the reader's background knowledge, impressions of fictional characters are created in the mind of the reader (cf. also Culpeper 2001). The project focuses in particular on the potential effects of patterns in speech and descriptions of body language. To be able to study such patterns in literary texts (especially in Dickens and other 19th century authors), we built a python module that extracts and annotates quotes (indicated by <qs/> and <qe/>) and suspensions (indicated by <sls/> and <sle/>), where a suspension is a narratorial interruption of a character's speech (see also Mahlberg & Smith 2012):

```
<qs/>'I am sure you will agree with
me, Ma,'<qe/> <sls/> said Mr.
Crisparkle, after thinking the matter
over, <sle/> <qs/>'that the first
thing to be done, is, to put these
young people as much at their ease as
possible.'<qe/>
```

2 Suspensions in DNov and 19C

Suspensions seem to be a characteristic feature of Dickens's style. Figure 1 shows how a corpus of Dickens's 15 novels (DNov) sets itself apart from other 19 century novels (19C): in Dickens the percentage of words in suspensions is higher than the reference corpus 19C. Figure 2 shows that suspensions are generally more frequent in DNov than in 19C.

3 Pilot study

This poster presents results of a pilot study exploring metrics to describe the inherent structure of the data, i.e. a corpus in which quotes, non-quotes, and suspensions are tagged. One of our research questions at the macro-level is: does Dickens's use of quotes and suspensions differ from that of other 19th century authors? Secondly, we explore the functions of suspensions and the way they relate to quotes in more detail. So at the micro-level we investigate: what are the potential effects that Dickensian suspensions can achieve?

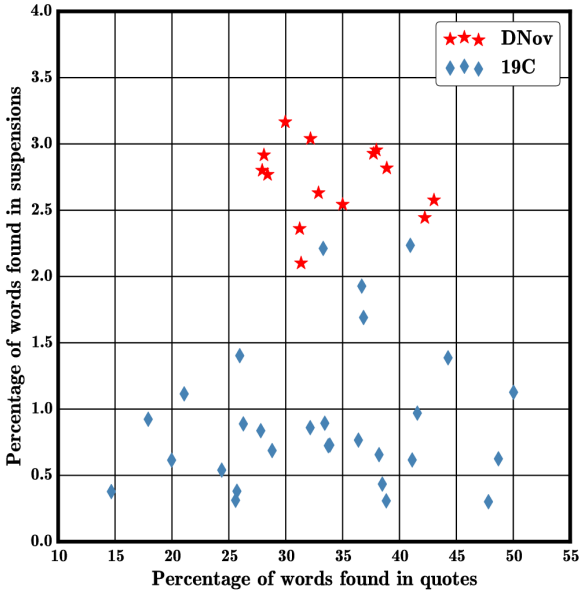


Figure 1. Quotes and Suspensions in DNov and 19C

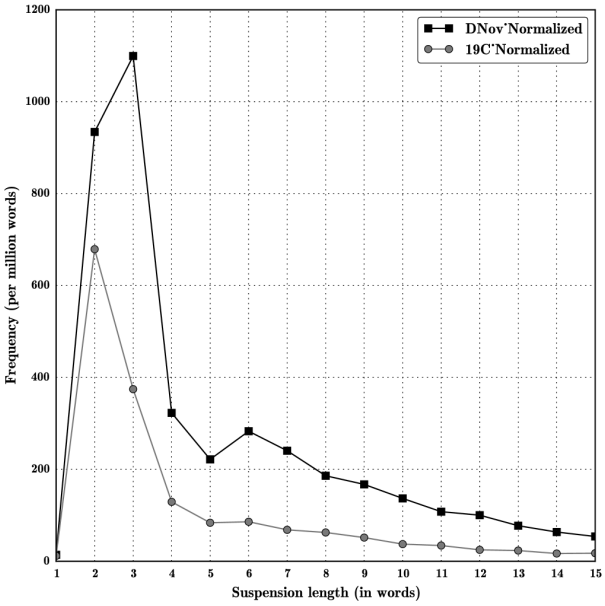


Figure 2. Suspension Length in DNov and 19C

The data we study include quote lengths, number of quotes, number of suspensions, lengths of suspensions, and suspension to quote ratios, and we are particularly interested in correlations between features in these data sets. One of our observations is illustrated by the scatterplot of average quote length over percentage of quotes in Figure 3, which suggests that Dickens prefers to give his characters shorter stretches of uninterrupted speech than other authors do.

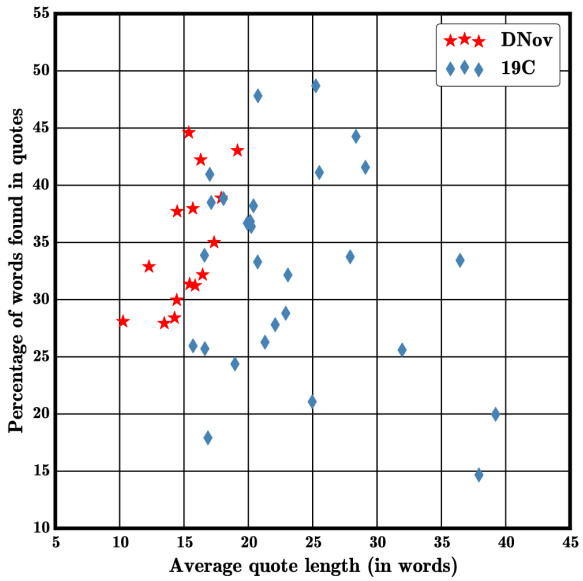


Figure 3. Quote Length in DNov and 19C

Figure 3 needs to be interpreted in relation to the properties of suspensions, so we specifically investigate *how* suspensions interrupt speech in DNov. Are there patterns in the length of the quote before it is interrupted? Do certain lengths trigger certain character descriptions, such as, for instance, body language presentations? What is the purpose of these interruptions and how might they affect the reader's mind-modelling processes? Our pilot study explores some of the lexical and structural properties of suspensions and outlines textual functions associated with them.

4 Work in progress

As part of the CLiC project, we also develop a user interface to support the systematic investigation of patterns in fictional speech and suspensions. The results of this pilot study will inform the development of this interface. CLiC 1.0 is freely accessible here: clic.nottingham.ac.uk, where we are also aiming to share the workflow used to annotate and explore our data.

Acknowledgement

The CLiC Dickens project is supported by the UK Arts and Humanities Research Council Grant Reference AH/K005146/1.

References

Mahlberg, M., & Smith, C. 2012. Dickens, the Suspended Quotation and the Corpus. *Language and Literature*, 21(1), 51–65. doi:10.1177/0963947011432058

Stockwell, P. 2002. *Cognitive Poetics an Introduction*. Hoboken: Routledge.

Culpeper, J. 2001. Language and Characterisation. People in Plays and Other Texts. Harlow: Pearson Education.

Comparing sentiment annotations in English, Italian and Russian

Marilena Di Bari
University of Leeds
mlmdb@leeds.ac.uk

1 Introduction

Corpus-based approaches to the study of the evaluative language have become more and more common, because of the greater availability of data and the advances in *Corpus Linguistics* techniques. At the same time, the challenges represented by the automatic identification and classification of the evaluative language have aroused increasing interest thanks to the advances in *sentiment analysis* (Liu 2010).

The link between these two disciplines is therefore undeniable, on one side because "by attempting to derive algorithms for identifying evaluative language automatically, sentiment analysis adds greatly to our understanding of what is important for evaluative meaning" (Hunston 2010) and, on the other side, because sentiment analysis needs to rely on a structured and functional study of the language, rather than treating it simply as a 'bag of words' (Harris 1954).

For such reason, some works have explored the possibility of using Martin and White's *Appraisal framework* (2005) not only for the identification of positive and negative opinions, but also of the categories of 'affect', 'appreciation' and 'judgement' (Taboada and Grieve 2004, Whitelaw et al. 2005, Argamon et al. 2009, Bloom and Argamon 2009).

The present work has addressed these two goals, by applying a specifically tailored annotation scheme called *SentiML* (Di Bari et al. 2013) in English, Italian and Russian. Because one of its most important features is the annotation of the contextual sentiment, this study aims at particularly pointing out the way in which sentiment dictionaries could be improved.

The paper will consist as follows: Section 2 briefly describes the annotation scheme, Section 3 presents the composition of the corpora in the three languages and some details about the annotation phase, Section 4 reports results.

2 SentiML annotation scheme

SentiML consists of three categories: **targets**, **modifiers** and **appraisal groups** (Di Bari et al. 2013).

A target is any entity (object, person or concept) that is implicitly or explicitly regarded as positive or

negative by the author of the text, e.g. "people". It usually consists of one word and has two attributes: *type* and *orientation*. *Type* captures the type of target and can be 'person', 'thing', 'place', 'action' or 'other'. *Orientation* captures the prior (out-of-context) orientation and can be 'positive', 'negative', 'neutral' or 'ambiguous'.

A modifier is what modifies the target, e.g. "good" It usually consists of one word and has four attributes: *orientation, attitude, polarity* and *force*. *Orientation* has the same values described for targets. *Attitude* refers to the *Appraisal Framework* (Martin and White 2005) and can be 'affect', 'appreciation' or 'judgement' depending on whether the target is the self, a thing or a person.

Polarity captures the presence of a negation and can be 'marked' or 'unmarked'. *Force* refers to the intensity of the modifier and can be 'high' (e.g. "very good"), 'low' (e.g. "not very good") 'reverse' (e.g. "at all good") or 'normal'.

An appraisal group represents an opinion on a specific target. For this reason, it is defined as the link between the target and the modifier, e.g. in "people are good":

[[People] $_{\text{TARGET}}$ are [good]$_{\text{MODIFIER}}$]$_{\text{APPRAISAL\_GROUP}}$

Appraisal groups have only the attribute *orientation*, which is actually the contextual one, as opposed to that annotated for targets and modifiers. It can be 'positive', 'negative', 'neutral' or 'ambiguous'.

3 Annotation and resulted corpora

SentiML has been applied to three different text types:

- **Political speeches**. American presidents' addresses [129] in English, and their translations in Italian[130] and Russian[131].
- **TED (Technology, Entertainment, Design) talks** in English, and their translations in Italian and Russian (Cettolo et al. 2012).
- **News**. Belonging to the *MPQA opinion corpus* (Wilson 2008) for English, to *Sole24ore* [132] for Italian and to *Project Syndicate* [133] and *Global Voices* [134] for Russian.

328 sentences have been annotated in Italian, 459 sentences in Russian and 462 sentences in English as completion of the previous phase. Annotations were revised when the first cycle was completed, and their errors analysed (Di Bari et al. 2014).

The annotation task was carried out by using MAE (Stubbs 2011), a freely available annotation environment.

4 Results

In Table 1 the amount of annotated categories according to language and text type is shown. It is interesting to notice that in all languages, political speeches are the richest in terms of appraisal groups (and thus sentiment), whereas news the poorest.

Lang	Type	Appraisal groups	Targets	Modifiers
EN	Political	624	519	551
	News	236	194	197
	TED	349	326	297
IT	Political	486	411	437
	News	254	203	244
	TED	341	292	323
RU	Political	599	510	542
	News	221	191	214
	TED	288	246	264

Table 1: Amount of annotated categories according to language and text type

In Table 2 all the values of orientation for each category are shown according to the language. In this case it is evident that positive opinions are more common than the negative ones in all three languages, followed by few ambiguous.

Lang	Category	Positive	Negative	Neutral	Ambiguous
EN	Appraisal groups	744	440	2	23
	Targets	165	200	477	215
	Modifiers	294	189	178	481
IT	Appraisal groups	723	345	0	13
	Targets	247	146	334	186
	Modifiers	299	141	143	467
RU	Appraisal groups	736	362	0	10
	Targets	284	124	400	151
	Modifiers	382	178	120	367

Table 2: Orientation summary for each category according to language

Table 3 shows the results of the comparison between the contextual orientation manually annotated by us, and the prior orientation included in the sentiment dictionaries. We used the *NRC Word-Emotion Association Lexicon* (Mohammad 2011) manually annotated in English, and its translation in Italian and Russian.

We calculated that words coming from the

[129] http://avalon.law.yale.edu/subject_menus/inaug.asp
[130] http://www.repubblica.it/2009/01/sezioni/esteri
[131] http://iipdigital.usembassy.gov/iipdigital-en/index.html
[132] http://www.ilsole24ore.com/
[133] https://www.project-syndicate.org/
[134] http://globalvoicesonline.org/

appraisal groups are present in the sentiment dictionary only in the following percentages: 35.07% in English, 30.11% in Italian and 10.29% in Russian. It is interesting to notice that some of those not included in the dictionaries are actually common, such as "difference", "dialogue", "example", "exercises".

Lang	Type of words	Frequency	Percentage
EN	Agreeing	590	69.57%
	Disagreeing	244	28.77%
	Ambiguous	14	1.65%
IT	Agreeing	454	69.63%
	Disagreeing	190	29.18%
	Ambiguous	7	1.07%
RU	Agreeing	152	66.67%
	Disagreeing	71	31.14%
	Ambiguous	5	2.19%

Table 3: Results of the comparison between prior and contextual orientation.

We decided to classify those included in the dictionaries in 3 categories:

- Agreeing words: words whose dictionary orientation agrees with that of the appraisal group they are taken from.
- Disagreeing words: words whose dictionary orientation does not agree with that of the appraisal group they are taken from.
- Ambiguous words: words who already have both positive and negative values in the dictionary.

In Table 3, for each language we show the number of times in which prior orientation and contextual orientation are agreeing, disagreeing and ambiguous for words taken from the appraisal groups and present in the sentiment dictionary.

Agreeing words cover between 66% and 69% of the total times words were found in the dictionary. The list generally includes reasonable out-of-context positive words (e.g. "love", "liberty", "leisure", "bless"), as well as out-of-context negative words (e.g. "criticism", "hypocrisy", "hostile", "blame"). This means that we can rely to a certain extent to the dictionary orientation, but not if we aim at more accuracy.

Disagreeing words cover between 28% and 31% of the total times words were found in the dictionary. This is important as it shows how crucial the context is, for example in the case of "maximum", "important", "demand", "balance".

Finally, ambiguous words are accounting only for 1-2%.

As for the other attributes, we found a difference in the most common *attitude*: in English it is 'judgement', straightly followed by 'appreciation', whereas in Italian and Russian it is 'appreciation'. In all of them the most common target *type* is 'thing'.
Also very interesting was to find out that the amount of marked *polarity,* i.e. presence of negation, and the order in the values of force (i.e. normal, high, reverse, low) is almost the same across the languages.

5 Conclusions

In this paper we have demonstrated that in all three languages the prior orientation given in the dictionary is different from the correct one given by the context: in English this happens in 28.77% of cases, in Italian in 29.18% and in Russian in 31.14%. In addition, the dictionaries have a relatively low coverage: 35.07% in English, 30.11% in Italian and 10.29% in Russian.
We have already worked on a complete sentiment analysis system based on the annotated data, and we aim at releasing our resources soon. In the meanwhile, the original and annotated texts, along with the Document Type Definition (DTD) to be used with MAE are already available[135].

References

Argamon, S., Bloom, K., Esuli, A. and Sebastiani, F. "Automatically Determining Attitude Type and Force for Sentiment Analysis" In Zygmunt Vetulani & Hans Uszkoreit (eds.) *Human Language Technology. Challenges of the Information Society*, Springer-Verlag: 218 – 231.

Bloom, K. and Argamon, S. 2009. "Automated learning of appraisal extraction patterns". *Language and Computers* 71: 249–260.

Cettolo, M., Girardi, C. and Federico, M. 2012. "Wit3: Web inventory of transcribed and translated talks". In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy.

Di Bari, M., Sharoff, S. and Thomas, M. 2014. "Multiple views as aid to linguistic annotation error analysis". In *Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII), ACL SIGANN Workshop held in conjunction with Coling 2014*, Dublin, Ireland. Available online at http://www.aclweb.org/anthology/W14-4912.

Di Bari M., Sharoff S. and Thomas M., 2013. "SentiML: functional annotation for multilingual sentiment analysis". In *Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: metadata, vocabularies and techniques in the Digital Humanities, held in conjunction with DocEng 2013*, Florence, Italy. Available online at http://dl.acm.org/citation.cfm?doid=2517978.2517994.

Harris, Z. 1954. *Distributional structure*. Word, 10, 146-

[135] http://corpus.leeds.ac.uk/marilena/SentiML

162.

Hunston, S. 2010. *Corpus approaches to evaluation Phraseology and Evaluative Language.* In Routledge Advances in Corpus Linguistics, Taylor & Francis.

Liu, B. 2012. *Sentiment Analysis and Opinion Mining.* Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Martin, J. R. and White, P. R. R. 2005. *The language of evaluation*. Basingstoke: Palgrave Macmillan Great Britain.

Mohammad, S. "From once upon a time to happily ever after: Tracking emotions in novels and fairy tales". *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Portland, USA.

Stubbs, A. 2011. "Mae and mai: Lightweight annotation and adjudication tools". *Linguistic Annotation Workshop*.

Taboada, M. and Grieve, J. 2004. "Analyzing Appraisal Automatically". *American Association for Artificial Intelligence Spring Symposium on Exploring Attitude and Affect in Text.* Stanford. AAAI Technical Report SS-04-07: 158-161.

Whitelaw, C., Navendu, G. and Argamon. S. 2005. "Using appraisal groups for sentiment analysis". *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM'05)*, Bremen, DE.

Wilson, T. A. 2008. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. PhD thesis, University of Pittsburgh.

Beauty and the Beast: The Terminology of Cosmetics in Romanian Dictionaries

Iulia Drăghici
University of Bucharest
alicadus@gmail.com

The fashion, cosmetics and plastic surgery industries have thrived on our century's preoccupation with physical appearance, all over the world. The fundamental social-economic changes of the past decades in Romania have led to a broad lexical and terminological dynamics. Alongside other terminologies (the terminology of Economics, the terminology of IT), the terminology of cosmetics (TC), a linguistic field of growing importance, puts the Romanian specialized cosmetic vocabulary in direct relationship with the common lexis and emphasizes the ability of the Romanian language to get richer through assimilated lexical loans (Frenchisms, Anglicisms, Italianisms) or by stimulating its lexical creativity (calques/loan translations, compounds, derived words etc).

Although of particular interest for its relationship with the common vocabulary, the terminology of cosmetics still lacks an elaborate scientific approach. It should be noted that, unlike other terminologies, the terminology of cosmetics cannot be found in contexts of academic level and specialized texts. TC began its existence as a sub-branch of Biology, Chemistry, Pharmacology etc., the first cosmetic terms of the Romanian language having originated in the above-mentioned related fields: *loţiune* (lotion), *masaj* (massage), *pomadă* (ointment), *săpun* (soap), *tratament* (treatment).

The growing interest of the large public for specialized terms from different fields drives a more careful analysis of specialized terminologies from the linguistic perspective. The study of terminologies is currently gaining ground as a consequence of their on-going, dynamic, relation with the common vocabulary which urges a less limited approach. As far as TC is concerned, we have noticed that a large series of specialized cosmetic terms (along with the ones "borrowed" from other terminologies), are continuously extending into the less specialized, common, communication in current Romanian, as a consequence of the "laicization" of knowledge, specific to the modern society. The media favours the expansion of these specialized cosmetic terms beyond the specialists' field (the *internal terminology*) and emphasizes their use with increased frequency in the common language (the *external terminology*).

Therefore, the need to properly understand specialized terms of ever wider circulation and their extension into various types of communication requires proper decoding as the natural result of proper defining. The definition of specialized terms must follow specific coordinates, regardless of their level of use. Our paper focuses mainly on the medium level of the cosmetic discourse. Our study of cosmetic terms' occurrences in texts of lower degree of specialization or even in common communication will take into consideration the correctness of non-specialist usage of the specialized meanings of cosmetic terms, highlighting any changes of meaning or any possible newly-emerged meanings.

The present paper is based on the finding that the terminology of cosmetics has not yet been studied systematically in Romanian linguistics. Our approach is mainly descriptive and tackles the dynamic semantics of the "old" TC terms (mainly Frenchisms) with respect to the meaning of their first "appearance" in Romanian dictionaries, on the one hand.

On the other hand, it analyzes the category of "absolute" novelties (*concealer*, *gloss, lipstick*, *peeling*, *smokey eyes* etc.) which includes Anglicisms still unrecorded in Romanian dictionaries, but whose (frequent) use is certified by the current general discourse of cosmetics and body care techniques. The corpus of the TC terms under investigation is taken from Romanian popular beauty catalogues and glossy magazines (*Avon*, *Bravo Girl*, *Cool Girl*, *Cosmopolitan*, *Glamour*) as well as from specialized training courses on cosmetics (*Curs de cosmetică profesională*, *Manual de cosmetică*).

Our study considers the lexicographic definitions for a number of cosmetic terms from the corpus from the perspective of the paradigmatic analysis, marking the frequency of occurrence for the denotative cosmetic meaning and the development of any connotative meanings. It also discusses the extent to which Romanian dictionaries achieve the disambiguation needed for proper communication to take place. It follows the evolution of the terms from their emergence as neologisms in the dictionaries used (CADE[136], DLRM[137], DEX1[138], DEX2[139], MDN[140], DCR3[141]), the relationship with the already existing synonyms in the language, the extension or reduction of their meaning etc. In some cases, the paradigmatic analysis was combined with the syntagmatic one, comparing the meanings that common words have developed in the new contexts of language (the lexicographic definitions were compared to the "actualizations" of the cosmetic terms in the texts analyzed). The study was conducted from the perspective of external, descriptive terminology, the purpose being to highlight the semantic dynamics manifested by cosmetic terms which migrate towards the common lexis, these being selected according to their frequency in texts of medium level specialization.

We consider that the investigation of the terminology of cosmetics in the Romanian language proves rewarding both for lexicology and semantics as well as for terminography and lexicography through highlighting new terms or new meanings, already validated by current usage. It is our hope for this research to be the starting point in the making of a mini-dictionary of cosmetic terms in current Romanian.

[136] Candrea, I.A. and Adamescu, Ghe. 1929-1931, *The Illustrated Enciclopedic Dictionary of the Romanian Language Today and Yesterday* (*Dicţionarul limbii române din trecut şi de astăzi*), Bucharest, Cartea Românească Publishing Press.
[137] Macrea, D. (coord.) 1958, *The Dictionary of Modern Romanian Language* (*Dicţionarul limbii române moderne*), Bucharest, Academiei Române Publishing Press.
[138] Coteanu, I. (coord.), Seche, M., Seche, L. 1975, *The Explanatory Dictionary of the Romanian Language* (*Dicţionar explicativ al limbii române*), Bucharest, Academiei RSR Publishing Press.
[139] Institutul de Lingvistică „Iorgu Iordan - Alexandru Rosetti" al Academiei Române 1996, *The Explanatory Dictionary of the Romanian Language* (*Dicţionarul explicativ al limbii române*), Bucharest, Univers Enciclopedic Publishing Press.
[140] Marcu, F. 2008, *The Big Dictionary of Neologisms* (*Marele dicţionar de neologisme*), Bucharest, SaeculumVizual Publishing Press.
[141] Dimitrescu, F. (coord.), Ciolan, Al., Lupu, C. 2013, *The Dictionary of Recent Words* (*Dicţionar de cuvinte recente*), 3rd edition, Bucharest, Logos Publishing Press.

A territory-wide project to introduce data-driven learning for research writing purposes

John Flowerdew
City University Hong Kong
enjohnf@cityu.edu.hk

1 Goal of the presentation

This presentation will provide a description of the beginning stages of a Hong Kong Government-funded project which aims to disseminate the use of corpus-assisted approaches to the development of research writing skills among Hong Kong language instructors, supervisors and research students. It is anticipated that the impact of the project will be to familiarise language educators PhD students and supervisors in Hong Kong with the data-driven learning approach to research writing for publication. English Centres and Departments will be in a position to develop training in this approach in their respective universities and PhD students will have greater success in achieving research publication, both in terms of quality and quantity. It is anticipated that the project will have a snowball effect which will lead to adoption of the approach more widely in Hong Kong. This presentation describes the rationale and planning for the project and aims for an interactive session which engages with the audience and elicits their ideas and opinions on the on-going development of this work in progress.

2 Background

Academic writing for research publication takes place around the globe, involving, according to a recent account, 5.5 million scholars, 2,000 publishers and 17,500 research/higher education institutions (Lillis and Curry 2010). Universities worldwide are striving to increase the quantity, quality and impact of their research publications. This endeavor applies to research students, as well as faculty members, with international publication increasingly becoming a requirement for graduation at PhD and even Master's degree level. For many advanced academic writers, however, English is not their first language and so they need additional help in developing their skills in writing for publication. However, the training support offered to such writers tends to be sporadic in most jurisdictions. This is specifically the case in Hong Kong, which is the focus for this presentation (Kwan, 2010).

3 Corpus linguistics and language pedagogy

The potential of corpus techniques for investigating patterns of language is well established. Corpus techniques can provide information about the behavior of words, multi-word phrases, grammatical patterns, semantic and pragmatic features, and textual properties. Such information and the procedures for obtaining it have been demonstrated to be of great significance for language pedagogy (e.g. Cheng, 2012; Flowerdew, 2009).

Applications of corpus linguistics to language learning and teaching may be direct or indirect (Flowerdew, 2009). A direct application would be where learners themselves work with corpora. The direct approach - commonly referred to as "data-driven learning" (Johns, 2002) - is where learners interact directly with a corpus on their computers or other devices, using an interface. In data-driven learning, learners are seen as "language detectives", seeking answers to questions that can be found by means of corpus queries. Learners are detectives, because they are required to identify and analyse the recurrent patterns to be found in the corpus output lines and make their own generalisations. They may do this by working directly with the computer and the corpus or using data print-outs.

There are a considerable number of reports in the literature of successful applications of data-driven learning in the teaching of advanced academic writing. To take just two examples, Bianchi and Pazzaglia (2007) created a corpus for psychology students consisting of experimental articles in that discipline, the task being that students should write a research article of their own, using the corpus as a resource. In a second example, taking this sort of procedure a stage even further, Lee and Swales (2006) had a heterogeneous group of graduate students who created their own corpora specific to their particular discipline. These corpora were used as a resource for working on the writing required on their higher degree programmes. The proposed project will use a similar, but refined, approach to that of Lee and Swales (2006).

Since Lee and Swales, there have been numerous further approaches to corpus-based advanced research writing reported in the literature, as has been demonstrated by Boulton (2012), in a review of 20 empirical studies on such applications of specific-purpose corpus-based pedagogy. Another notable study (Davies, 2013) describes the potential of the use of the academic component of one particular on-line corpus – the Corpus of American English (COCA) – which will be employed as a data source in the proposed project.

4 Aim and Objectives of the Project

The overall aim of the project which will be reported on at *Corpus Linguistics 2015* is to introduce Hong Kong language educators, PhD students and supervisors (across all disciplines) to the benefits of using a data-driven learning approach to developing research students' competence in research writing for publication.

To achieve this aim, the specific objectives of the project are to:

1. create a small team to disseminate the data-driven learning approach to research writing
2. investigate to what extent, if any, language educators in Hong Kong are already familiar with, and implementing, the data-driven learning approach to research writing
3. train the project team in the data-driven learning approach to research writing
4. develop two training packages for data-driven learning for research writing: one for language educators and the other for PhD students and supervisors
5. disseminate the data-driven learning approach in a research writing context to language educators across university English centres and departments
6. train PhD students (across the disciplines) to use the data-driven learning approach
7. create one or more YouTube videos to show users how to use the data-driven approach to research writing
8. create an email list to assist users of the data-driven learning approach
9. develop a package to evaluate the effectiveness of the project

5 Content of the presentation

As previously stated, the presentation at *Corpus Linguistics 2015* will describe progress so far in the initial stages of the project and encourage interaction with the audience with the goal of eliciting ideas to inform the project further. At the time of writing, the project team are only beginning to work together, but by the time of the conference, it is anticipated that a certain amount of progress will be able to be reported, as the project will have progressed somewhat. Although a work in progress, it is hoped that the presentation will be of value in exchanging ideas about the potential of corpus techniques in the teaching and learning or research writing for publication.

References

Bianchi, F., & Pazzaglia, R. (2007). Student writing of research articles in a foreign language: Metacognition and corpora. In R. Facchinetti (Ed.), *Corpus linguistics 25 years on* (pp. 259-287). New York: Rodopi.

Boulton, A. (2012). *Corpus-informed research and learning in ESP: Issues and applications*. In A. Boulton, S. Carter-Thomas & E. Rowley-Jolivet (Eds.) *Corpus consultation for ESP: A review of empirical research.* (pp. 261-292). Amsterdam: John Benjamins.

Cheng, W. (2012). *Exploring corpus linguistics: Language in action.* London: Routledge

Davies, M. (2013). Google Scholar and COCA-Academic: Two very different approaches to examining academic English. *Journal of English for Academic Purposes*, 12(3), 155-165.

Flowerdew, J. (2009). Corpora in language teaching. In Long, M. H. & Doughty, C.J. (Eds.). *The handbook of language teaching*. (pp. 327-350). Oxford: Wiley-Blackwell.

Have you developed *your entrepreneurial skills*? Looking back to the development of a skills-oriented Higher Education

Maria Fotiadou
University of Sunderland
maria.fotiadou
@research.sunderland.ac.uk

1 Introduction and Background

The colonization of academia by the market is a topic that has been widely discussed and criticised especially within academia. According to Tomlinson (2005: 2), education in the UK was forced to become a 'competitive enterprise and a commodity, rather than a preparation for a democratic society' due to the government's 'fragmentation of social welfare via the introduction of market principles'. Universities around the world experienced substantial changes and a one-way route towards a free-market and a corporate-business perspective.

As a result, we are now talking about 'entrepreneurial universities' (Mautner 2005). Educational institutions have to function 'as if they were ordinary businesses competing to sell their products to customers' (Fairclough 1993: 143), and courses are not the only products on offer. It is being promoted that HE can offer the best route towards employment in this highly competitive and insecure job-market. The need to secure a job has become a prerequisite for both students and universities and this requirement is particularly emphasised with the global economic crisis in the forefront.

The government's solution to this problematic situation was, and still is, the creation and promotion of a skills-oriented education (Blair 1998b: 9 in Tomlinson 2005: 7). The belief that individuals are the only ones accountable for securing their future and well-being is commonly accepted and has also become 'naturalized' (Fairclough 2015) within HE as students are taught, through the Universities' careers services, that they are responsible to widen their marketable skills if they wish to survive in the competitive job market.

2 Critical Discourse Analysis and Corpus Linguistics

The analysis is based in Fairclough's three-layered model of Critical Discourse Analysis as the data need to be examined in their social context. On the other hand, this kind of analysis would not be possible without the application of Corpus Linguistics methods. It is a fact that CL methods help 'the analysis tackle research questions in ways that other methods cannot' (Mautner 2009: 44).

3 The World Wide Web as a source for building corpora

The World Wide Web is used as the main source of data collection for this project because it does indeed represent 'a treasure trove for building corpora that reflect current social developments much better than static corpora ever can' (Mautner 2009: 36).

The data were collected from ten Russell Group websites and more specifically from their *Careers and Employability* web pages. The *Wayback Machine*, which is a digital library with snapshots of the World Wide Web since 1996, was used to add a comparative and diachronic approach to the analysis.

Three corpora were built: Corpus A (138,000 words), Corpus B (235,000 words) and Corpus C (897,000 words). Corpora A and B were built using texts from the RG Careers web pages as documented in the digital library, from the years 2000 and 2007 respectively, while Corpus C is consisted of texts that are currently available on the internet.

4 Example

	Corpus A (2000)	Corpus B (2007)	Corpus C (2015)
1	careers	careers	careers
2	students	students	students
3	work	information	**your**
4	information	career	career
5	employer	**your**	graduate
6	service	service	skills
7	discipline	work	information
8	employers	skills	work
9	graduate	graduates	experience
10	application	graduate	university

Table 1- Keyword Analysis: Corpus A, B & C

As shown in Table 1, the keyword analysis highlights the rise in the use of the second person possessive determiner 'your'. In Corpus A, the same determiner is located at rank 30. This shows that, in more recent years, there has been an effort in pushing the weight of responsibility to the students. It shows that students are expected to invest in the development of their skills.

Since the aim of this project is to unravel the changes in language use towards a more neoliberal reality, I focus on 'skills' and more specifically on the concordances that follow the pattern *your * skills*, as they describe, differentiate and present a plethora of skills students are expected to develop while at university.

5 Discussion

It has been shown by various researchers that Critical Discourse Analysis and Corpus Linguistics methods 'can cooperate fruitfully and with mutual gain, building on shared interest in how language 'works' in social rather than merely structural terms' (Mautner 2009: 33). This project will combine both qualitative and quantitative methods, using the World Wide Web as its main source of data, in order to examine a problematic phenomenon.

References

Fairclough, N. (1993) 'Critical Discourse Analysis and the Marketization of Public Discourse: The Universities', *Discourse & Society*, 4(2), pp. 133–168.

Fairclough, N. (2015) *Language and power*. 3rd edn. Oxon: Routledge.

Internet Archive: Wayback Machine (no date). Available at: http://archive.org/web/ (Accessed: 15 January 2015).

Mautner, G. (2005) 'The Entrepreneurial University: A discursive profile of a higher education buzzword', *Critical Discourse Studies*, 2(2), pp. 95–120.

Mautner, G. (2009) 'Corpora and Critical Discourse Analysis', in Baker, P. (ed.) *Contemporary corpus linguistics*. London ; New York: Continuum, pp. 32–46.

Tomlinson, S. (2005) *Education in a post-welfare society*. 2nd edn. Maidenhead: Open University Press

Promoting Proficiency in Abstract Writing: A Corpus-Driven Study in Health Sciences

Ana Luiza Freitas
Federal University of Rio Grande do Sul, Brazil
alf@via-rs.net

Maria José Finatto
Federal University of Rio Grande do Sul, Brazil
mariafinatto@gmail.com

The topic of this presentation is the teaching abstract writing in academic genres in the field of health sciences in Brazilian universities, of which this presentation aims at sharing a pilot study. The research explores aspects of frequency, specificity and context of use of lexical items, and aims at developing a virtual learning environment as a form of systematization and socialization of the investigation findings. The corpus comprises abstracts in the areas of Pharmacy, Medicine and Nutrition coming from different universities in the aforementioned context, as well as published texts from three international journals. From the cross-checking of the texts an analysis and description of the recurring lexical patterns is to be produced so as to qualify corpus-driven academic teaching in Brazil. The principle in the investigation is the one according to which written proficiency in English for Specific Purposes (ESP, henceforth) is a construct present in texts which reflect a fluent discursive production to their field of knowledge. The supporting thesis is that in order to effectively accomplish the goals of international discursive proficiency Brazilian ESP language learners need to be able to produce fluent abstracts in which each move and lexical item in the rhetorical structure represent a purposeful part and convey a completeness of meaning. In other words, it seems reasonable to argue that learning how to properly combine words to build up texts in a particular written genre can aid higher education students in making appropriate language choices, and thereby in developing references about the type of language that they are expected to come up with in their academic writings. That way academic additional language users may become fully fledged members of that discourse community which shares the same class of events referred to by Swales (2004), and their forms of written register can be effective tools for their participation in the world of science, and in the arenas of specialized knowledge. As such, the study proposes to account for the following research questions: What kind of lexical variability is identified along the genres which make up the

research corpus? What are the fixed and the variable elements like? Which are the most frequent lexical unit combinations in the corpus and how do they behave structurally? The methodological fields adopted are the ones of Corpus Linguistics, ESP and Natural Languages Processing. The concept of language adopted is the one of a cognitive, historical and social activity (BIBER, 1988; SWALES, 1990; BAKHTIN, 1997) according to which the academic writing endeavour is an interactive verbal activity, geared towards the social actors in a communicative enterprise. Academic writing is defined as a project shaped over time by a group of participants engaged in a community that establishes guidelines for its functioning and discourse (SWALES, 1990, 1993). Furthermore, the academic text is conceived of as a form of cognitive activity which represents a completeness of of meaning through which specialized knowledge is expressed (HOFFMANN, 2004). Additionally, teaching and learning in such context aim at promoting competence for effective communication in the English speaking academic world. Inasmuch, this investigation also associates with the principle according to which linguistic traits do not happen randomly and language is guided by higher standards than just words (SINCLAIR, 1991). The pilot study (25,000 tokens) contrasted Brazilian abstracts with those published in international journals in the target field. The Antconc Software (LAWRENCE, 2011) was adopted as analytical tool, mainly through the use of the n-grams/clusters feature and occurrences of four written words were searched for in order to identify recurring patterns of lexical unit combinations. The outcomes suggested a tendency for a more frequent repetition of keywords, as well as an adoption of more characteristically academic language style and a higher use of Passive Voice constructions in the international corpus. Such preliminary conclusions should be further investigated in an expanded corpus (100,000 thousand tokens), as the thesis proposes to accomplish. Should they be confirmed through the cross-checking of the complete corpus though, these findings already point out some rich material to be systematised at the virtual learning environment. Furthermore, a classification approach based on similarity and syntax was built up to deal with the lexical unit combinations, which seemed to be a methodologically meaningful finding in itself, once it aided in highlighting aspects which would not possibly have been noticed otherwise.

The comparative study of the image of national minorities living in Central Europe

Milena Hebal-Jezierska
University of Warsaw
milena.hebal-jezierska@uw.edu.pl

The main objective of the talk is to present the images of chosen national minorities living in Czech Republic, Slovak Republic and Poland against of the background of possibilities of the corpora of West-slavic languages. The study provides comparative and contrastive perspective . It is carried out on three corpora (subcorpora of press): the corpus of the Czech language, the corpus of the Slovak language and the corpus of the Polish language. The selection of three cultural and linguistic areas allows to ensure objective conclusions. It seems equally important that the corpora of the Czech and Slovak languages differs significantly from the corpus of Polish in terms of the technique used in software which supports the corpora. This fact allows also to facilitate the drawing of objective conclusions concerning the relation of the corpus research methods with the technical possibilities of the corpus.

Firstly we are going to present the reconstruction of the auto-images (self percepion) of Czechs (Czechs: the small nation, ateists, Shvejks), Slovaks (brothers, under threat from Czechs and Hungarians, catholics) and Poles (drinkers, victims of persecution, heroes, catolics). After then hetero-images (perception of one group by another) of national minorities will be presented from the point of view of the Czechs, Slovaks and Poles. The list includes the following nationalities: Vietnamese, Ukrajinian, Russian, Hungarian, Polish, Slovak, Czech.

Due to the enormous material devoted to the Roma (this is material for a separate project), our study does not include the corpus research on Roma.

Auto-images, and hetero-images of mentioned nations will be created in two ways: descriptive and graphic. Graphically, we want to show both theradial network of a) forming a particular category of nationality (it consists of the characteristics of the nation, putting at the center and periphery) and b) creating a particular feature category (feature, in which nationality is the most severe). We will also show a relation between auto-image features to features of other nations - assuming that the perception of foreign population is manifested by the relation to the own characteristics (on the axis of "we-they" / "ours-stranger").

Two main corpus methods will be applied in our study: corpus-driven and corpus-based. By the corpus-driven method we mean the analysis of all corpus data without adjusting them to categories known from the non-corpus studies. By the corpus-based method we mean the use of corpora to verify the linguistic theory, in our case, check the specific elements of the image. This means that we test in the corpus the names of the nationality known from the non-corpus material For the auto-image and hetero-image we attempt to use the following corpus research methods: keywords, collocation profiles, including an contextual analysis (inclusion of collocation into broader context), pattern grammar (study pattern, in which occur certain lexemes occur), the list of words derived from the frequency distribution, lock words analysis, and to complete data analysis a random sample of occurences. The corpus-driven methods are completed by corpus-based methods.

They will be also subjected to critical analysis covering the following issues:

factors affecting the ability of the using the corpus research method (type of lexeme, type of corpus, technical token type, technical feasibility of the corpus),

advantages and disadvantages of using particular corpus research method. The worst image of the national minorities is portrayed on Czech press. It may be corelated with the central Czech auto-image feature "we are the small nation".

Very interesting results were obtained in reconstructing national minorities living in the Czech Republic. The Corpus of Czech language containing 300 million words was used to create the image of a foreigner, the Ukrainian and the Vietnamese. All these nations are seen mainly, but not only, in terms of the 'criminal', with the Vietnamese additionally as a 'seller/trader' (usually selling illegally). Studies conducted so far have shown that the same national communities are perceived differently by the nations which are so close to each other. An example is the case of the Vietnamese. In the Czech corpus, the words appears most often in the category of saler – criminals; in Slovak – man involved in war; and in Polish 'the Vietnamese' is associated with cuisine, trade, or war (cf. Hebal-Jezierska, M. 2011). Another example, the words Polish and Pole create in the corpus of Czech language categories as very conservative people, catholics, victims of percecution, the big nation, enterprising people; in the Slovak language they appear mainly in categories: enterprising people and tourists. It is worth noting that the Slovaks are portrayed in Polish press as people living in mountains.

References

Bańko, M., Doliński, I., Duda, J., Hebal-Jezierska, M., Collocation Images of Hungarians in Slavonic Languages, Practical Applications of Linguistic Research. In. A. Obrębska (ed.) Practical Applications of Linguistic Research. Łódź.

Baker, P. 2010. Sociolinguistics and Corpus Linguistics. Edinburgh.

Baker, P, Gabrielatos, C, Khosravinik, M, Krzyżanowski, M, McEnery, T., Wodak, R. 2008. „A Useful Methodological Synergy? Combining Critical Discourse Analysis and Corpus Linguistics to Examine Discourses of Refugees and Asylum Seekers in the UK Press". Discourse & Society 19(3): 273-305.Smith, X. 2003. "Some thoughts on submitting abstracts to conferences". In J. Jones and F. Farmer (eds.) All about conferences. London: Example Press.

Bartmiński, J. (ed.) 1999. Językowy obraz świata. Lublin. Smith, X. 2003. "Some thoughts on submitting abstracts to conferences". In J. Jones and F. Farmer (eds.) All about conferences. London: Example Press.

Duszak, A., Fairclough, N. 2008. Krytyczna analiza dyskursu. Interdyscyplinarne podejście do komunikacji społecznej. Kraków: Universitas.

Chlewiński, Z., Kurcz, I. 1992. Stereotypy i uprzedzenia. Warszawa.

Český národní korpus, www.korpus.cz

Čermák, F., Šulc, M. 1996. Kolokace. Praha.

Gabrielatos, C., Baker, P. 2008. „Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996-2005". Journal of English Linguistics 36(1), 5-38.

Glynn, D., Fischer, K. 2010. Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches. Berlin/NewYork

Hebal-Jezierska, M. 2011. Kolokační obrazy některých lexémů patřících do sémantického pole cizinec v českém tisku (s metodologickými úvahami). In. F. Čermák (eds.) Korpusová lingvistika Praha 2011, InterCorp. Praha, 109-121.

Hebal-Jezierska, M. 2012. The image of a lexeme based on the analysis of collocations. In. P. Pęzik (ed.) Corpus Data across Languages and Disciplines. Peter Lang, Frankurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Warszawa, Wien, 183-192.

Hebal-Jezierska, M. 2012. Wizerunki kolokacyjne mniejszości narodowych żyjących w Republice Czeskiej i Republice Słowacji. A talk from the conference Tertium: Słowo w kontekście. Kraków.

Haan, H., Scholz, S., Stereotyp, Identität und Geschichte: die Funktion von Stereotypen in gesellchaftlichen Diskursen.

Hunston, S., Francis, G. 2000. Pattern Grammar, Amsterdam/Philadephia.

Kaderka, P. 2002. Etnické kategorizování v médiích. "Vesmír" 81, č. 5, p. 247-248.

Lešnerová, Š., Uhle, D., Wojda, A., Gorzkowski, A. (eds.) 2002. Obraz vzájemných vztahů Čechů, Poláků a Němců v jejich jazycích, literaturách a kulturách. FF UK : Praha.

McEnery, T, Wilson, A. 1996. Corpus Linguistics. Edinburgh.

Narodowy Korpus Języka Polskiego: www. nkjp.pl

Orłowski, H. ed. 2005. Stereotype in interkultureller Wahrnehmung. Nysa.

Partington, A.,1998. Patterns and Meanings, Amsteredam/Philadephia.

Partington, A.,2006. The Linguistics of Laughter. London/New York.

Petranová, D., Plencner, A. 2008. Stereotypy v médiách. Trnava.

Narodowy Korpus Języka Polskiego, www.nkjp.pl

Slovenský národný korpus, www.korpus.sk

Stefanowitsch, A., Gries S. 2006. Corpus-Based Approaches to Metaphor and Metonymy. Berlin/New York.

Sutaj, S., ed. 2004. Narod a narodnosti na Slovensku. Stav vyskumu po roku 1989 a jeho perspektivy. Presov.

Tognini-Bonelli, E. 2001. Corpus Linguistics at Work. Amsterdam.

Wodak, R., Krzyżanowski, M. 2011. Jakościowa analiza dyskursu w naukach społecznych. Warszawa: Oficyna Wydawnicza Łośgraf.

Investigating discourse markers in spontaneous embodied interactions: Multi-modal corpus-based approach

Kazuki Hata
Newcastle University
k.hata@ncl.ac.uk

The present study is designed to analyse the use of discourse markers (DMs) and gestures-in-talk, namely symbolic movements semantically and pragmatically accompanying with spoken expressions (McNeill 1992; Kendon 2004). In the literature, DMs have been explored by a number of studies for over the past few decades, clarifying their forms and significance in talk (see Östman 1982; Schiffrin 1987; Fraser 1990; Redeker 1991; Maschler 1994; Knott 1995; Brinton 1996; Romero Trillo 1997; Hansen 1998; Schourup 1999; Oates 2000; Andersen 2001; Carter and McCarthy 2006). However, investigations of DMs have been predominantly carried out using text-based linguistic analyses (see Thompson 2005; Knight 2011a), and therefore arguably overlook that the nature of human interaction is multi-modal whereby participants deliver the message in both speech and kinesic movements (i.e. gestures) (Birdwhistell 1970; Allen 1999: 470; Richmond & McCroskey 1999: 2). In fact, it has been claimed that not only spoken words but also co-expressed gestures contribute to organising discourse, highlighting the implication of possible correlations between DMs and gestures (see Bavelas 1994; Kendon 1995; McNeill & Pedelty 1995; Quek et al. 2002; Adolphs & Carter 2007; Adolphs & Knight 2008; Ferré 2011; Knight 2011b). Therefore, the gap arguably needs to be bridged in order to fully explore the discourse-marking functions of DMs in embodied interactions and the correlation between linguistic items and co-expressed gestures.

Given this, the research questions are twofold:

1) Are there any statistical patterns between specific spoken DMs and semiotic gesture types (see McNeill 1992)?
2) How do spoken DMs and co-expressed gestures contribute to managing discourse by their communicative functions?

These questions will be explored to highlight the significance of discourse-marking functions in embodied interactions and to refine the use of DMs from the multi-modal view.

For research purpose, the study utilises a multi-modal corpus-based methodology to investigate the use of spoken DMs and co-expressed gestures; thus, video-recording devices are involved to capture

speakers' embodied behaviours (Allwood 2008; Knight 2011a). From the quantitative perspective, a statistical corpus-based approach will be applied to generate a frequency list of DMs and co-expressed gestures; this process highlights the statistical patterns by comparing single spoken DMs and those with gestures. This statistic analysis also contributes to making the focus for further micro-level investigations from long-term recordings: for instance, Conversation Analysis and Discourse Analysis (Walsh & O'Keeffe 2010: 142). Then, the study examines communicative functions of them in depth, refining correlations between DMs and accompanying gestures; this process examines how their communicative functions are related.

The project have firstly handled approximately the two hour-long group discussion of the topic 'Educational Psychology', retrieved from the Newcastle University Corpus of Academic Spoken English (NUCASE). As the pilot attempt, the presented paper handled the first ten-minute short excerpt of the data and found some significant implications for the correlations between DMs and co-expressed gestures. For example, sequencing DMs (e.g. *but*, *and*, *so*), which signal a sequential relationship between segments of talk at the textual level (Fraser 1990; Schiffrin 1987; Brinton 1996; Rouchota 1996; Carter & McCarthy 2006), were often seen to be accompanied with co-expressed gestures which arguably demonstrate the speaker's attitude toward the basic message and/or signal the focus on the upcoming message as the significant part of the talk (see Cassell, McNeill & McCullough 1999: 5; McNeill 1992: 15); this will be the case beyond 'message-based relationships across sentence' (Schiffrin 2001: 67; see also Halliday 1971, 1979; Östman, 1981, 1982; Blakemore 1987, 1989, 2002). From this point, it is assumed that the co-expressed gestures will contribute to DM's multi-functionality which have been debated very much in the literature (see Schiffrin 2001; Aijmer & Simon-Vandenbergen, 2006; Redeker 2006; Aijmer 2013).

The proposed research questions what roles DMs and gestures play and how they are related. This is potentially a platform for analysing DMs and accompanying gestures utilising multi-modal corpus-based approach. The project is in its very earliest phases and thus has investigated only the short excerpt of the entire data. Nevertheless, this pilot study already highlights some implications regarding the correlations between spoken DMs and co-expressed gestures. As I discovered that only ten minute-short excerpt generates concrete findings, it is rational to make further attempts to investigate the entire data. Thus, the rest of the data will be transcribed and annotated for conducting 1) a statistical analysis and 2) fine-grained multi-modal investigations.

References

Adolphs, S. and Carter, R. 2007. "Beyond the word: New challenges in analysing corpora of spoken English". *European Journal of English Studies* 11(2): 133–146.

Adolphs, S. and Knight, D. 2008. "Analysing Discourse Markers: A Multi-Modal Approach". In the *British Association for Applied Linguistics Annual Conference (BAAL 2008)*, University of Swansea.

Aijmer, K. 2013. "Analysing modal adverbs as modal particles and discourse markers". In L. Degand, B. Cornillie, & P. Pietrandrea (eds.) *Discourse markers and modal particles: Categorization and description*. Amsterdam: John Benjamins.

Aijmer, K. & Simon-Vandenbergen, A. 2006. *Pragmatic markers in contrast*. Amsterdam: Elsevier.

Allen, L. Q. 1999. "Functions of nonverbal communication in teaching and learning a foreign language". *The French Review* 72 (3): 469–480.

Allwood, J. 2008. "Multimodal Corpora". In A. Lüdeling, and M. Kytö (eds) *Corpus Linguistics: An international handbook*. Berlin: Mouton de Gruyter.

Andersen, G. 2001. *Pragmatic markers and sociolinguistic variation: A relevance-theoretic approach to the language of adolescents*. Amsterdam: John Benjamins.

Bavelas, J. B. 1994. "Gestures as part of speech: Methodological implications". *Research on Language & Social Interaction* 27 (3): 201–221.

Blakemore, D. 1987. *Semantic Constraints on Relevance*. Oxford: Blackwell.

Blakemore, D. 1989. "Denial and contrast: A relevance theoretic analysis of *but*". *Linguistics and philosophy* 12(1): 15–37.

Blakemore, D. 2002. *Relevance and linguistic meaning: The semantics and pragmatics of discourse markers.* Cambridge: Cambridge University Press.

Birdwhistell, R. L. 1970. *Kinesics and context: Essays on body motion communication*. Philadelphia: University of Pennsylvania Press.

Brinton, L. J. 1996. *Pragmatic markers in English: Grammaticalization and discourse functions*. Berlin: Walter de Gruyter.

Carter, R. and McCarthy, M. 2006. *Cambridge grammar of English: a comprehensive guide: spoken and written English grammar and usage*. Cambridge: Cambridge University Press.

Cassell, J., McNeill, D. and McCullough, K. 1999. "Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information". *Pragmatics & Cognition* 7(1): 1–33.

Ferré, G. 2011. "Multimodal analysis of discourse markers 'donc', 'alors' and 'en fait' in conversational French". In *Actes de ICPhS XVII*: 671–674.

Fraser, B. 1990. "An approach to discourse markers". *Journal of Pragmatics* 14: 383–395.

Halliday, M. A. K. 1970. “Language structure and language function”. *New Horizons in Linguistics* 1: 140–165.

Halliday, M. A. K. 1979. “Modes of meaning and modes of expression: Types of grammatical structure and their determination by different semantic functions”. In D. Allerton, E. Carney, D. Hollcroft (eds.) *Functions and context in linguistic analysis*. Cambridge: Cambridge University Press.

Hansen, M. B. M. (1998). “The semantic status of discourse markers”. *Lingua* 104 (3): 235–260.

Kendon, A. 1995. “Gestures as illocutionary and discourse structure markers in Southern Italian conversation”. *Journal of Pragmatics* 23 (3): 247–279.

Kendon, A. 2004. *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.

Knight, D. 2011a. “The future of multimodal corpora”. *Revista Brasileira de Linguística Aplicada* 11 (2): 391–415.

Knight, D. 2011b. *Multimodality and active listenership: A corpus approach*. London: Continuum.

Knott, A. 1995. *A data-driven methodology for motivating a set of coherence relations*. Unpublished Ph.D. thesis, University of Edinburgh.

Kurtić, E. et al. 2012. “A corpus of spontaneous multi-party conversation in Bosnian Serbo-Croatian and British English”. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*: 1323–1327.

Labov, W. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. 1997. “Some further steps in narrative analysis”. *Journal of Narrative and Life History* 7: 395–415.

Maschler, Y. 1994. “Metalanguaging and discourse markers in bilingual conversation”. *Language in Society* 23 (3): 325–366.

McNeill, D. 1992. *Hand and mind: What gestures reveal about thought*. London: University of Chicago Press.

McNeill, D. and Pedelty, L. 1995. “Right brain and gesture”. In K. Emmorey & J. S. Reilly (eds.) *Language, gesture, and space*. New Jersey: Lawrence Erlbaum Associates.

Oates, S. L. 2000. “Multiple discourse marker occurrence: Creating hierarchies for natural language generation”. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*: 41–45.

Östman, J. O. 1981. *‘You know’: A discourse-functional study*. Amsterdam: John Benjamins.

Östman, J. O. 1982. “The symbiotic relationship between pragmatic particles and impromptu speech”. In N. E. Enkvist (ed.) *Impromptu Speech: A Symposium. Papers Contributed to a Symposium on Problems in the Linguistic Study of Impromptu Speech*: 147–177.

Quek, F. et al. 2002. “Multimodal human discourse: gesture and speech”. *ACM Transactions on Computer-Human Interaction* 9 (3): 171–193.

Redeker, G. 1991. “Review article: linguistic markers of discourse structure”. *Linguistics* 29: 1139–1172.

Redeker, G. 2006. “Discourse markers as attentional cues at discourse transitions”. In K. Fischer (ed.) *Approaches to discourse particles*. Amsterdam: Elsevier.

Richmond, V. P. and McCroskey, J. C. 1999. *Nonverbal behavior in interpersonal relations*. 4th ed. Boston: Allyn and Bacon.

Romero Trillo, J. 1997. “Your attention, please: Pragmatic mechanisms to obtain the addressee's attention in English and Spanish conversations”. *Journal of Pragmatics* 28 (2): 205–221.

Rouchota, V. 1996. “Discourse connectives: What do they link?”. *UCL Working Papers in Linguistics* 8: 51–65.

Schiffrin, D. 1987. *Discourse markers*. Cambridge: Cambridge University Press.

Schiffrin, D. 2001. “Discourse markers: Language, meaning, and context”. In D. Schiffrin, D. Tannen & H. E. Hamilton (eds.) *The Handbook of Discourse Analysis*. Oxford: Blackwell.

Schourup, L. 1999. “Discourse markers”. *Lingua* 107 (3): 227–265.

Thompson, P. 2005. “Spoken language corpora”. In M. Wynne (ed.) *Developing linguistic corpora: A guide to good practice*. Oxford: Oxbow Books.

Walsh, S. and O’Keeffe, A. 2010. “Investigating higher education seminar talk”. *Novitas-ROYAL (Research on Youth and Language)* 4 (2): 141–158.

A resource for the diachronic study of scientific English: Introducing the Royal Society Corpus

Ashraf Khamis
Saarland University
ashraf.khamis@uni-saarland.de

Stefania Degaetano-Ortlieb
Saarland University
s.degaetano@mx.uni-saarland.de

Hannah Kermes
Saarland University
h.kermes@mx.uni-saarland.de

Jörg Knappen
Saarland University
j.knappen@mx.uni-saarland.de

Noam Ordan
Saarland University
noam.ordan@uni-saarland.de

Elke Teich
Saarland University
e.teich@mx.uni-saarland.de

There is a wealth of corpus resources for the study of contemporary scientific English, ranging from written vs. spoken mode to expert vs. learner productions as well as different genres, registers and domains (e.g. MICASE (Simpson et al. 2002), BAWE (Nesi 2011) and SciTex (Degaetano-Ortlieb et al. 2013)). The multi-genre corpora of English (notably BNC and COCA) include fair amounts of scientific text too.

Diachronic resources of scientific texts are more limited in that existing corpora are typically fairly small, including only few small samples per discipline (e.g. ARCHER with approximately 258,000 words covering all scientific disciplines in British and American English texts (Biber et al. 1994) and the Coruña Corpus in which 10,000 words are taken to represent astronomy in the 18th and 19th centuries (Moskowich and Crespo 2007)) or covering one discipline only (e.g. the corpus of Early Modern English Medical Texts (Taavitsainen et al. 2011)).

To increase the pool of corpus resources for the diachronic study of scientific English, we are building a corpus from the Philosophical Transactions and Proceedings of the Royal Society of London, starting from the date of their inception (1665) to modern time. At present, we work on processing materials from the period 1776 to 1869 (2,454 articles amounting to around 23 million tokens), with other periods to follow. The materials contain texts from a variety of scientific areas ranging from biology, chemistry, physics and geography to medicine.

We describe the steps we take to get from the source materials to a usable corpus, focusing in particular on the interaction of automatic and manual processing. The source materials are in XML format and contain metadata on journal, title, author and year of publication. Although the texts are partially structured, they need a considerable amount of preprocessing, including cleaning of OCR errors and hidden markup, ordering of scrambled pages, identification of article beginnings and endings and removal of duplicates, headers and footers. After preprocessing, we normalize the texts using VARD (Baron and Rayson 2008), annotate them for tokens, lemmas and parts-of-speech using TreeTagger (Schmid 1994) and finally encode the corpus in Corpus Query Processor (CQP) format (Evert and Hardie 2011). Furthermore, we mark up document structure as provided by the XML source as well as century, fifty-year period and decade so as to enable analyses on different temporal resolution frames.

Once a reasonable level of data quality has been reached, the Royal Society Corpus will be made available through CLARIN-D. In our own research, we use the corpus to study the diachronic development of scientific English as a distinct discourse type as well as register diversification, applying various methods of data mining.

References

Baron, A. and Rayson, P. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. *Postgraduate Conference in Corpus Linguistics* 2008, May 22. Birmingham, UK: Aston University.

Biber, D., Finegan, E. and Atkinson, D. 1994. ARCHER and its challenges: Compiling and exploring A Representative Corpus of Historical English Registers. In U. Fries, P. Schneider and G. Tottie (eds.), *Creating and using English language corpora*, 1–14. Amsterdam/New York: Rodopi.

Degaetano-Ortlieb, S., Kermes, H., Lapshinova-Koltunski, E. and Teich, E. 2013. SciTex - a diachronic corpus for analyzing the development of scientific registers. In P. Bennett, M. Durrell, S. Scheible and R. J. Whitt (eds.), *New methods in historical corpus linguistics: Corpus linguistics and interdisciplinary perspectives on language (CLIP)*, vol. 3. Tübingen: Narr.

Evert, S. and Hardie, A. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*. Birmingham, UK.

Moskowich, I. and Crespo, B. 2007. Presenting the Coruña Corpus: A collection of samples for the historical study of English scientific writing. In J. Pérez-Guerra, D. González-Álvarez, J. L. Bueno Alonso and E. Rama-Martínez (eds.), *'Of varying language and opposing creed': New insights into Late Modern English*, 341–357. Bern: Peter Lang.

Nesi, H. 2011. BAWE: An introduction to a new resource. In A. Frankenberg-Garcia, L. Flowerdew and

G. Aston (eds.), *New trends in corpora and language learning*, 213–228. London: Continuum.

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, 44–49. Manchester, UK.

Simpson, R. C., Briggs, S. L., Ovens, J. and Swales, J. M. 2002. *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

Taavitsainen, I., Jones, P. M., Pahta, P., Hiltunen, T., Marttila, V., Ratia, M., Suhr, C. and Tyrkkö, J. 2011. Medical texts in 1500–1700 and the corpus of Early Modern English Medical Texts. In I. Taavitsainen and P. Pahta (eds.), *Medical writing in Early Modern English*, 9–29. Cambridge: Cambridge University Press.

SYN2015: a representative corpus of contemporary written Czech

Michal Křen
Institute of the Czech National Corpus
Charles University
michal.kren@ff.cuni.cz

1 Background

The Czech National Corpus aims at extensive and continuous mapping of the Czech language and its varieties. This effort results in compilation, maintenance and providing access to a number of corpora (synchronic/diachronic, written/spoken etc.), including corpora of contemporary written Czech making up the SYN series.[142]

The SYN-series corpora can be described as traditional (as opposed to the web-crawled corpora), featuring cleared copyright issues, well-defined composition, reliability of annotation and high-quality text processing (Hnátková et al. 2014). All the corpora are also disjoint, i.e. any document can be included only into one of them.

2 Representative corpora of the SYN series

Currently, the SYN series consists of three large newspaper corpora with total size exceeding 2 billion tokens and three 100-million corpora representative of written Czech (SYN2000, SYN2005, and SYN2010; the number denotes the corpus publication year).

The representative corpora cover three consecutive time periods in a regular five-year interval (i.e. SYN2010 covers the 2005–2009 period) and they contain a large variety of written genres in proportions based on language reception studies (Králík and Šulc 2005). Their design, strengths and weaknesses are described in detail in Křen (2013: 46–53) including the comparability, which is desirable to enable modern diachronic studies.

3 Design of SYN2015

The aim of this paper is to introduce SYN2015, a 100-million corpus of contemporary Czech. SYN2015 will be a continuation of the series, but at the same time, it will reflect necessary methodological and technical changes outlined below.

SYN2015 is designed as a representation of the printed language of 2010–2014. Specific language of the internet (discussion forums, blogs etc.) is kept

[142] http://ucnk.ff.cuni.cz/english/struktura.php

separately and it will be covered by a newly-established NET corpus series.

The original text classification scheme of the SYN series has been updated and revised. The revised classification is also based on external criteria and it is designed with maximum compatibility with the original scheme in mind. This means that changes have been made only where necessary; the most significant enhancements are sub-classification of professional texts adopted from the National library and more detailed classification of newspaper texts including separate annotation of sections wherever possible.

In line with its predecessors, SYN2015 will contain a large variety of texts from various publishers within the given classification category. Proportions of the particular categories in SYN2015 will be set arbitrarily (i.e. the corpus will not be claimed balanced), yet close to the original figures. The proportions will be fixed and observed also in future representative corpora of the series. For instance, the three top-level categories of fiction / professional literature / newspapers and magazines will share one third of the corpus each. This approach emphasizes representation of a language by covering its variability and corresponds to the Biber's notion of representativeness in terms of "texts as products" (Biber 1993:245).

SYN2015 will be supplemented by further enhanced search interface KonText[143] which enables users to examine corpus composition and to make use of the wide variety of included texts intuitively and effectively.

4 Technical enhancements

Tools used for processing the SYN-series corpora combine fully automatic steps (foreign languages detection, de-duplication etc.) with human-supervised and even manual ones (text classification interface). This is necessary to keep high quality standards that are not compromised despite the growing amount of the data.

However, most of the tools have been in use for more than 10 years and are thus already outdated. This is why the whole toolchain has been completely rebuilt using standard and up-to-date tools that fully support XML and UTF8. The update includes also a major enhancement of the tokenization, lemmatization and POS-tagging module (Hnátková et al. 2104). As a result, the data processing should be much easier and faster while retaining the present quality.

[143] Corpus query interface developed as an enhancement of the NoSketch Engine and based on Manatee as the backend (Rychlý 2007; Machálek and Křen 2013); KonText is available at http://kontext.korpus.cz/

5 Conclusion

SYN2015 is currently in preparation and it will be released by the end of 2015 within the framework of the Czech National Corpus.[144]

Acknowledgement

The corpus design, compilation and annotation are a result of team work carried out during the implementation of the Czech National Corpus project (LM2011023) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

References

Biber, D. 1993. "Representativeness in Corpus Design". *Literary and Linguistic Computing* 8 (4): 243–257.

Hnátková, M., Křen, M., Procházka, P. and Skoumalová, H. 2014. "The SYN-series corpora of written Czech". In *Proceedings of LREC 2014*. Reykjavík: ELRA, 160–164. Available online at http://www.lrec-conf.org/proceedings/lrec2014/pdf/294_Paper.pdf

Králík, J. and Šulc, M. 2005. "The Representativeness of Czech Corpora". *International Journal of Corpus Linguistics* 10 (3): 357–366.

Křen, M. 2013. *Odraz jazykových změn v synchronních korpusech*. Prague: NLN.

Machálek, T. and Křen, M. 2013. "Query interface for diverse corpus types". In *Natural Language Processing, Corpus Linguistics, E-learning*. Lüdenscheid: RAM Verlag, 166–173.

Rychlý, P. 2007. "Manatee/Bonito - A Modular Corpus Manager." In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University, 65–70.

[144] http://www.korpus.cz/

Adversarial strategies in the 2012 US presidential election debates

Camille Laporte
University of Leeds
encgl@leeds.ac.uk

Adversarial relations in political discourse seldom occur the way they do in electoral debates, when two leaders go face to face, in front of wide national (and sometimes international) audiences. I am considering here how such adversarial relations occur in the 2012 United States (U.S) Presidential election debate series between the Democratic party candidates (Barack Obama and Joe Biden) and the Republican party candidates (Mitt Romney and Paul Ryan).

The purpose of this analysis is threefold. I first consider how questions and answers participate in the building-up of adversarial relations between the candidates. Second, how and when are rhetorical questions used and to what effect in relation to expressing adversarial relations in this context? Finally, I review the role played by non-verbal means of communication in terms of displaying adversarial relationships in the confrontational context of these four debates.

This study is informed by the perspective of Brown and Levinson's theory of politeness, which presupposes a system of "face threatening" and "face management" (Brown & Levinson, 1987: 24). It is based on a 70, 000 words subcorpus, extracted from my purpose-built 2.7 million word corpus of political discourse from the UK, US, and France. Using Wordsmith tools, (Scott, 2005) I have extracted the relevant data, and transcribed it using methods derived from Clayman and Heritage (2002) in order to provide indications on conversation analysis.

The adversarial relations studied here are found in different types of interactions, questions and answers being among the most prominent, that is, questions from the moderators to the candidates, but also rhetorical questions as a means of responding to moderators' questions. In relation to this, I study Clayman's three "forms of pressure" available through questions (2010: 265-268); setting the agenda, incorporating presuppositions (Clayman, 2010: 266), and questions designed to invite a certain type of answer (such as yes/no questions, negative interrogatives and question prefaces).

I also refer to Archer's review of how relationships of power are created through questions and answers (Archer, 2005: 16), quoting the work of Spencer Oatey (1992: 108) on the three different types this categorisation includes: coercion, expertise and legitimacy (Spencer-Oatey, 1992: 108).

In addition, I consider non-verbal communication utilised by the candidates in this debate series, in order to provide a comprehensive analysis of how adversarial relations are created in this context.

References

Archer, D. 2005. *Questions and answers in the English courtroom (1640-1760): a sociopragmatic analysis*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Brown, P. & Levinson, S. 1978. "Universals in language usage: politeness phenomena". In E. N. Goody, (ed), *Questions and Politeness: Strategies in Social Interaction*. Cambridge: Cambridge University Press.

Clayman, S. & Heritage, J. 2002. *The News Interview. Journalistic and Public Figures on the Air*. Cambridge: Cambridge University Press.

Clayman, S. 2010. Questions in Broadcast Journalism. In Freed, A. F. & Ehrlich, S. (eds) *"Why Do You Ask" The function of Questions in Institutional Discourse.* New York: Oxford University Press, pp. 256-278.

Spencer-Oatey, H. 1992. *Cross-Cultural Politeness: British and Chinese Conceptions of the Tutor-Student relationship*. Unpublished Ph.D. thesis. Lancaster University.

Structuring a CMC corpus of political tweets in TEI: corpus features, ethics and workflow

Julien Longhi
Université de Cergy-Pontoise - CRTF
Julien.Longhi@u-cergy.fr

Ciara R. Wigham
Université Lumière Lyon 2 - ICAR
ciara.wigham@univ-lyon2.fr

The CoMeRe project (CoMeRe, 2014) aims to build a kernel corpus of computer-mediated communication (CMC) genres with interactions in the French language. Three key words characterize the project: variety, standards and openness. The project gathered mono- and multimodal, synchronous and asynchronous communication data from both Internet and telecommunication networks (text chat, tweets, SMSs, forums, blogs). A variety of interactions was sought: public or private interactions as well as interactions from informal, learning and professional situations.

Whereas some CMC data types were collected within the CoMeRe project, others had previously been collected and structured within different project partners' local research teams. This meant that the project had to overcome disparities in corpus compilation choices. For this reason, the CoMeRe project structured the corpora in a uniform way using the Text Encoding Initiative format (TEI, Burnard & Bauman, 2013) and decided to describe each corpus using Dublin Core and OLAC standards for metadata (DCMI, 2014; OLAC, 2008). The TEI model was extended in order to encompass the Interaction Space (IS) of CMC multimodal discourse (Chanier *et al.,* 2014).

The term 'openness' also characterizes the project: The corpora have been released as open data on the French national platform of linguistic resources (ORTOLANG, 2013) in order to pave the way for scientific examination by partners not involved in the project as well as replicative and culumative research.

This poster presentation aims to give an overview of the corpus building process using, as a case study, a corpus of political tweets *cmr-polititweets* (Longhi *et al.,* 2014). The corpus stemmed from a local research project on lexicon (Digital Humanities and datajournalism, supported by the Fondation of Cergy-Pontoise University). It was built starting from seven French politicians from six different political parties. In order to generate political tweets, a set of lists citing these politicians was generated (7087 lists), and lists that have tweeted at least six times and for which the description contained the word 'politics' were selected (120 lists in total). Finally, 2934 tweets were recovered. In order to be sure that we selected politicians' tweets (and not, for example, those of journalists), only the accounts cited in more than 12 lists were considered; 205 politicians were tweeting. We took the last 200 tweets of each of the 205 accounts on 27 March 2014 (34,273 tweets). This allowed us to recover data that focused on the period between the two rounds of the 2014 municipal elections in France.

The poster will focus, firstly, on how features specific to Twitter were included and structured in the interaction space TEI model. We will exemplify how features including *hashtags* that label tweets so that other users can see tweets on the same topic, *at signs* that allow a user to mention or reply to other users and *retweets* that allow a user to repost a message from another Twitter user and share it with his own followers, were integrated into the model. Secondly, the poster will evoke some of the ethical and rights issues that had to be considered before publishing a corpus of tweets. Finally, the workflow & multi-stage quality control process adopted during the building of the corpus will be illustrated. This was an essential aspect considering that the corpus underwent format conversions: the local research team had initially structured the corpus in XML whilst the CoMeRe project applied the IS TEI model to the corpus.

The political tweets corpus is now structured and available online. Analyses have started to be carried out: some ideas have been launched in Djemili *et al.* (2014) but further analyses must adhere rigorously to methodologies stemming from the natural language processing (NLP) field.

References

CoMeRe Repository (2014). Repository for the CoMeRe corpora [website], http://hdl.handle.net/11403/comere

Burnard, L. & Bauman, S. (2013). TEI P5: Guidelines for electronic text encoding and interchange. TEI consortium, *tei-c.org*. http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C.R., Hriba, L., Longhi, J. & Seddah, D. (2014). The CoMeRe corpus for French: structuring and annotation heterogeneous CMC genres, in Beißwenger, M., Oostdijk, N., Storrer, A & van den Heuvel, H. Building and Annotating Corpora of Computer-Mediated Discourse: Issues and Challenges at the Interface of Corpus and Computational Linguistics, *Journal of Language Technology and Computational Linguistics* (special issue). pp1-31. http://www.jlcl.org/2014_Heft2/Heft2-2014.pdf

Djemili S., Longhi J., Marinica C., Kotzinos D. & Sarfati G.-E. (2014). « What does Twitter have to say about ideology », Konvens 2014 - *Workshop proceedings vol. 1* (NLP 4 CMC: Natural Language Processing for

Computer-Mediated Communication / Social Media – Pre-conference workshop at Konvens2014) , Germany (2014), p.16-25.

DCMI (2014). Dublin Core Metadata Initiative. http://dublincore.org/

Longhi, J., Marinica, C., Borzic, B. & Alkhouli, A. (2014). Polititweets, corpus de tweets provenant de comptes politiques influents. In Chanier T. (ed) *Banque de corpus CoMeRe*. Ortolang.fr : Nancy. http://hdl.handle.net/11403/comere/cmr-polititweets

OLAC. (2008). Best Practice Recommendations for Language Resource Description. *Open Language Archives Community*. University of Pennsylvania. http://www.languagearchives.org/REC/bpr.html

ORTOLANG (2013). Open Resources and TOols for LANGuage [website]. ATILF / CNRS - Université de Lorraine: Nancy, http://www.ortolang.fr

Patterns of parliamentary discourse during critical events: the example of anti-terrorist legislation

Rebecca McKee
University of Manchester
rebecca.mckee@postgrad.manchester.ac.uk

This corpus linguistic study looks at the political discourse of MPs, specifically at differences between the representation of ethnic minorities by ethnic minority and non-ethnic minority MPs. This is conducted by looking at text from the Hansard records from the debates on Anti-Terrorism legislation, critical junctures where it was important that the voice of UK ethnic minorities be represented.

The growth of far right political parties, coupled with increasing immigration and more ethnically heterogeneous societies, has highlighted the need to understand how the particular interests of ethnic minorities are being represented by Members of Parliament in the political processes. This study analyses the situation in the United Kingdom parliament, which in 2010 saw the election of a record number of MPs from ethnic minority backgrounds. On these grounds it would appear that there is at least some degree of descriptive representation of these groups, whereby these MPs share certain characteristics of ethnicity, religion and culture with these minority populations. What is less clear is whether this translates into substantive representation.

Taking Hanna Pitkin's (1967) argument, these MPs may not necessarily act on behalf of or in the interests of those that they represent descriptively. These MPs may seek instead to advance other policy preferences and interests. However, Philips (1995) has argued that, simply by their presence, they increase the probability of substantive representation. This study, which is part of a larger one on political representation of ethnic minorities in the UK, draws on Jane Mansbridge's (1999) argument that, even though those who descriptively represent ethnic minorities may not do so all of the time, they are more likely to at times of "critical events" that call for the specific views of minorities to be represented.

This study examines a series of such critical events, the passage of successive anti-terrorist legislation in the UK. Some of this legislation has been tabled in reaction to individual events, such as the 9/11 and 7/7 attacks in the USA and London; others reflect more general concerns, now invariably focused on the perceived threat from Islamic

fundamentalism. However, many of these laws have been criticised as being discriminatory and encouraging a more general Islamophobia. Thus it is important that MPs represent the interests of ethnic minorities as this legislation have been shown to adversely affect these communities, including stop and search powers.

There is some anecdotal evidence that ethnic minority MPs are aware of their role as descriptive representatives as Keith Vaz has asked in Parliament that MPs "send out a clear message to [minority communities] that they are on our side and we are on their side in dealing with those elements who seek to subvert our democracy" (HC Deb, 15th February 2006, c1448) whilst Ashok Kumar stated "I recently met members of the Hindu community who expressed their concerns that they have not been involved in consultation and their concerns about the legislation" (HC Deb, 13 February 2006, c1124). This provides a springboard from which to further investigate the role of ethnic minority MPs, whether there are differences between them and non-minority MPs and thus some support for the theory that there is a link between descriptive and substantive political representation of ethnic minorities.

The data source for this study is Hansard, the official UK parliamentary record, specifically the passage of six anti-terrorist laws, from the 2001 Terrorism, Crime and Security Act to the 2014 Counter-terrorism and Security Bill. The Hansard records have been used successful in corpus linguistics analysis before (Baker 2004, 2009) and this study takes inspiration from this use of corpus linguistics to analyse questions more routed in social sciences and other studies (Baker et al. 2013). Unlike previous studies, this includes not only debates on the floor of the house but also committee proceedings.

The material was converted into text files compatible with WordSmith and separated into files of speech from ethnic minority and non-minority MPs. The approach to analysis corpora involved corpus linguistics methods including keyword analysis. The analysis compares the speech from the two groups of MPs, with 15 ethnic minority MPs contributing to any one of the discussions out of a possible 37 ethnic minority MPs in this time period. The corpus of ethnic minority MPs speech (42,735 words) is compared to the corpus of non-minority MPs (>1.2 million words).

In line with critical events theory there is an expectation that results will show that there is a difference in the speech between the two groups of MPs and that the ethnic minority MPs will be more in line with the interests of those that they descriptively represent and more mindful of protecting these interests in the context of the anti-terror legislation.

References

Baker, Paul. 2004. "'Unnatural acts' Discourses of homosexuality within the House of Lords debates on gay male law reform." *Journal of Sociolinguistics* 8 (1): 88-106.

Baker, Paul. 2009. "'The question is, how cruel is it?' Keywords, Fox Hunting and the House of Commons." In *What's in a Word-list? Investigating word frequency and keyword extraction*, edited by Dawn Archer, 125-36. Surrey: Ashgate Publishing Ltd.

Baker, Paul, Costas Gabrielatos, and Tony McEnery. 2013. *Discourse analysis and media attitudes:the representation of Islam in the British press.*Cambridge; New York: Cambridge University Press.

Mansbridge, Jane. 1999. "Should Blacks Represent Blacks and Women Represent Women? A Contingent 'Yes'." *The Journal of Politics* 61 (3): 628-57.

Phillips, Anne. 1995. *The Politics of Presencee*. New York: Oxford University Press.

Pitkin, Hanna Fenichel. 1967. *The concept of representation*: Berkeley: University of California Press.

A Linguistic Analysis of NEST and NNEST Employer Constructs: An Exploratory Multi-method Study

Corrie B. MacMillan
St. John's University
macmilla@mail.sju.edu.tw

1 Rationale

The global spread of English has led to the deconstruction of the native English speaker (NES), and non-native English speaker (NNES) identity as problematic (Davies, 2003; Graddol, 2003; Moussu & Llurda, 2008; Phillipson, 1992). Moreover, the notion that the native speaker (NS) is the ideal language teacher has been critiqued as a 'myth' (Davies, 2003) and a 'fallacy' (Phillipson, 1992, 2007). Yet, my role as an English language teacher in Taiwan was defined by my NES status. Many of my peers, both local and foreign, confirm these observations with their own shared experiences.

This led me to conduct a pilot study of ELT employer preferences in the EFL context of Taiwan for my MA dissertation. The purpose of which was to explore the concept of 'native-speakerism' (Holliday, 2006) in a significant EFL context and to examine if the hiring practices in Taiwan can be reasonably framed as being discriminatory in nature (Holliday, 2006; Selvi, 2010, 2011). The pilot study indicated that foreign teachers in the private EFL sector of Taiwan are primarily valued as NESs, not NESTs. Yet the limitations of this study indicate that further research is necessary to construct a sufficient NES profile, as well as the reasons for these preferences.

2 Literature

Three studies were found which address the issue of discriminatory hiring practices with surveys (see Clark & Paran, 2007; Mahboob, Uhrg, Newman, & Hartford, 2004; Medgyes, 2001). All three studies reported findings from Inner Circle contexts (Kachru, 1997). All three studies indicated employer preferences for NESs.

The abundant quantifiable data potentials of online job posts have not been analysed in a significant amount of studies. Only three articles were found which made any reference to such data. Moussu and Llurda (2008) refer to thousands of positions advertising in several different contexts supporting Native-speakerism, which is evidenced with a footnote directing the reader to Dave's ESL Café (http://www.eslcafe/joblist). Beckett and Stiefvater (2009) analyze a single job post for a position in China from The Linguist List (www.linguitlist.org). Selvi (2010) does collect and analyse a significant amount of data: however, rather than provide a representative sample of any single context, the study generalizes about the entire ELT field based on data collected from two English websites: Dave's ESL Café and TESOL (www.tesol.org). A corpus-based method was not applied and the study is hard to replicate as no clear description of the method of analysis is provided. A corpus-driven, cross-linguistic analysis of the specific and significant EFL context of Taiwan should contribute to the empirical evidence exploring Native-speakerism.

3 Research Questions

- How does the language in ELT job posts and English language school advertisements construct discourses of NESs and NNESs in the EFL context of Taiwan?
- How does this discourse contrast with the established theoretical constructs of NSs and NNSs within Applied Linguistics?

4 Design of Study

The proposed study aims to improve upon the limitations of the pilot study conducted as my MA dissertation. The intention is to carry out a multi-method study which applies a discourse analysis of corpora (Baker, 2006) as the primary method of investigation. As triangulation methods, surveys and interviews will be conducted to properly contextualize the corpus findings.

5 Significance of Results

Based on indications from the pilot study, successful completion of the proposed research should contribute to the limited empirical evidence exploring employer preferences regarding NESTs and NNESTs. The multi-method cross-linguistic study should indicate ELT employer preferences in the Mandarin-Chinese EFL contexts of Taiwan. If properly contextualized, a corpus-driven discourse analysis supported by the triangulation methods (surveys and interviews) are presumed to evidence preferences for foreign teachers valued as NESs/NESTs and local Chinese English teachers valued as NNESTs. Moreover, it is presumed that the NES is valued primarily as a model of the target English language. This would indicate the field of ELT in Taiwan has not moved beyond the Native Speaker (Cook, 1999). It would also indicate to what extent the theoretical construct of the Native Speaker is represented in practice.

References

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London ; New York: Continuum.

Beckett, G. H., & Stiefvater, A. (2009). Change in ESL graduate students' perspectives on non-native English-speaker teachers. *TESL Canada Journal*, 27(1), 27-46.

Clark, E., & Paran, A. (2007). The employability of non-native-speaker teachers of EFL: *A UK survey. System*, 35(4), 407-430.

Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly*, 33(2), 185-209.

Davies, A. (2003). *The Native Speaker: Myth and Reality.* Clevedon: Multilingual Matters Ltd.

Graddol, D. (2003). The decline of the native speaker. In G. Anderman & M. Rogers (Eds.), T*ranslation Today: Trends and Perspectives* (pp. 152-167). Clevedon: Multilingual Matters Ltd.

Holliday, A. (2006). Native-speakerism. *ELT journal*, 60(4), 385-387.

Kachru, B. B. (1997). World Englishes and English-using communities. *Annual Review of Applied Linguistics*, 17, 66-87.

Mahboob, A., Uhrg, K., Newman, K. L., & Hartford, B. S. (2004). Children of a lesser English: status of nonnative English speakers as college-level English as a second language teachers in the United States. In L. D. Kamhi-Stein (Ed.), *Learning and Teaching from Experience: Perspectives on Nonnative English-Speaking Professionals* (pp. 100-120). Ann Arbor: The University of Michigan Press.

Medgyes, P. (2001). When the teacher is a non-native speaker. In M. Celce-Murcia (Ed.), *Teaching English as a Second or Foreign Language* (pp. 429-442). Boston: Heinle & Heinle.

Moussu, L., & Llurda, E. (2008). Non-native English-speaking English language teachers: History and research. *Language Teaching,* 41(03), 315-348.

Phillipson, R. (1992). *Linguistic Imperialism*. Oxford: Oxford University Press.

Phillipson, R. (2007). Linguistic imperialism: a conspiracy, or a conspiracy of silence? *Language policy*, 6(3), 377-383.

Selvi, A. F. (2010). All teachers are equal, but some teachers are more equal than others: trend analysis of job advertisements in English language teaching. WATESOL NNEST *Caucus Annual Review*, 1, 156-181.

Selvi, A. F. (2011). The non-native speaker teacher. *ELT journal*, 65(2), 187-189

Textual patterns and fictional worlds: Comparing the linguistic depiction of the African natives in *Heart of Darkness* and in two Italian translations

Lorenzo Mastropierro
University of Nottingham
lorenzo.mastropierro
@nottingham.ac.uk

1 Introduction

This paper focuses on the fictional representation of the African natives in Joseph Conrad's *Heart of Darkness* (1899) and in two of its Italian translations. It adopts a corpus stylistic approach to investigate recurrent lexico-semantic patterns that not only participate directly in constructing this aspect of the fictional world, but also play an important role for the critical interpretation of the text. In translation, alterations in these lexico-semantic patterns might affect the fictional representation of the African natives. In turn, this altered fictional representation might trigger a different reception of this aspect for the target reader. Therefore, this paper adopts corpus methods to compare two Italian translations of *Heart of Darkness* on the basis of the patterns identified in the original, in order to study to what extent alterations are made in the translations and whether these alterations affect the text reception.

The interaction between translation studies and corpus linguistics has been recently at the centre of much research interest (for example Kruger et al. 2013 or Oakes and Ji 2012). This analysis is at the forefront of these recent developments and aims to contribute to this vibrant and multidisciplinary field.

2 Methodology

The analysis focuses on five words used in the short novel to refer to the Africans: *nigger*(*s*), *negro*, *savage*(*s*), *black*(*s*), and *native*(*s*). These terms, the 'native words', are studied as core items of 'functionally complete units of meaning' (Tognini-Bonelli 2001). Particular attention is given to the notions of semantic preference and semantic prosody, as their analysis is particularly effective in the study of short texts such as *Heart of Darkness*. In fact, investigating dominant semantic fields makes it possible to "group together lower frequency words and multiword expressions which would, by themselves, not be identified as key, and would otherwise be overlooked" (Rayson 2008: 544). This allows the present study to account for shared and

cumulative effects created by low-frequency items. The identification of semantic preferences and prosodies points to the motifs reoccurring with the 'native words', as well as further evaluative meanings assigned to them. This methodology recalls Gabrielatos and Baker's (2008) study of semantic preference and semantic prosody as a means to create specific *topoi* around a given item and to embed it with attitudinal meanings.

The second part of the analysis involves the comparison of the two Italian target texts. Having investigated the 'native words' as functionally complete units of meaning allows the comparison to be based on functional equivalence (Tognini-Bonelli 2001). As such, the analysis looks at how the 'native words' have been translated and examines whether they reproduce or not the same textual behaviour and function identified in the original. Particular attention is given to the effects of using different 'native words' in translation, as well as analysing what happens to the semantic preference and prosody when the original lexico-semantic patterns are altered.

3 Analysis

Race and the depiction of the African natives in *Heart of Darkness* are central concerns in many critical studies of Conrad's work (for example Achebe 1990; Hawkins 2006; Lawtoo 2012). The present analysis combines this critical discussion with the linguistic perspective of the corpus approach in order to contribute to the understanding of such a major theme of the text, i.e. the fictional representation of the natives. In particular, looking at the 'native words' and their textual behaviour, the analysis aims to examine how the lexical level of the text constructs and reflects this theme. It is argued that the lexico-semantic patterns identified create and maintain a dehumanising tendency in the way the natives are depicted, a tendency that finds confirmation in critical interpretations. This seems to indicate a connection between the patterns and the critical reading of the text.

The analysis of the two Italian target texts aims to study the effects of translating on the connection between the linguistic and the interpretational level in *Heart of Darkness*. Alterations to the linguistic level might result in alterations to the reading of the target text, with a potential mismatch between the reception of the original and the reception of the translation. For example, the choice of different 'native words' from the original or the modification of the patterns that construct the dehumanising tendency can generate these discrepancies which, in turn, have the potential to signal the translator's agenda behind the translation choices.

4 Conclusion

Looking at how formal patterns can convey literary meaning, this paper contributes to the study of the relation between lexis and major themes in literary texts. It argues that lexico-semantic patterns act as building blocks of the fictional world (Mahlberg 2013) and as such can play a role in the text interpretation. Consequentially, they are of great relevance in the context of literary translation too, where the preservation of the text form is as important as the maintenance of its meaning. Therefore, this paper also shows the effects of the translation practice on the link between lexis and major themes in literature. It argues that alterations of the original textual features can potentially affect the reading of the translated text and thus manipulate its reception.

Finally, this paper provides an example of how corpus methods can be applied to the study of translation, specifically to literary translation, contributing to the development of the interaction between the two fields.

References

Achebe, C. 1990. "An image of Africa: Racism in Conrad's *Heart of Darkness*". In C. Achebe (ed.) *Hopes and Impediments: Selected Essays*. New York: Anchor Books.

Conrad, J. 1899. *Heart of Darknes*. *Blackwood's Edinburgh Magazine* CLXV February-April.

Gabrielatos, C. and Baker, P. 2008. "Fleeing, sneaking, flooding: A corpus analysis of discoursive constructions of refugees and asylum seekers in the UK press, 1996-2005". *Journal of English Linguistics* 36 (1): 5-38.

Hawkins, H. 2006. "*Heart of Darkness* and racism", in P. B. Armstrong (ed.) *Heart of Darkness*. New York/London: Norton.

Kruger, A., Wallmach, K. and Munday, J. (eds.) 2013. *Corpus-based Translation Studies: Research and Applications*. London: Bloomsbury.

Lawtoo, N. 2012. "Conrad's *Heart of Darkness* and Contemporary Thought: Revisiting the Horror with Lacoue-Labarthe". Huntingdon: Bloomsbury Publishing.

Mahlberg, M. 2013. *Corpus Stylistics and Dickens's Fiction*. London and New York: Routledge.

Oakes, M. P. and Ji, M. (eds.) 2012. *Quantitative Methods in Corpus-based Translation Studies*. Amsterdam: John Benjamins.

Rayson, P. 2008. "From key words to key semantic domains". *International Journal of Corpus Linguistics* 13 (4): 519-549.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Relating a Corpus of Educational Materials to the Common European Framework of Reference

Mícheál J. Ó Meachair
Trinity College, Dublin
michealomeachair@gmail.com

1 Introduction

While reviewing the literature relating to corpora and educational materials, I have identified a bias in pedagogical corpus-linguistic research. The majority of the research appears to focus on research *for* education materials (FEM) and fewer pieces of research focus on research *on* educational materials (OEM). This is interesting because research has found intermittent or no use of corpora by creators of educational-materials. Several reasons are given for this, including lack of familiarity with corpora and insufficient computer skills (Burton, 2012). One could also explain this underdevelopment, as Meunier (2002: 123) did, by recognizing that "learner corpora research is still in its infancy". While some people can be shown how to use corpora to inform their educational materials, it appears as though this will not be an option in every case; particularly where appropriate corpora are not known or not available, as is the case for some minority or small-community languages. Should corpus analyses *on* educational materials be more widely conducted, we may build on the valuable materials currently in existence, rather than starting from scratch.

In conducting this study I will illustrate some research methods available to corpus-linguists who seek to evaluate written text currently being used in education (therefore conducting OEM research), with the aim of relating the materials to the CEFR. This paper take will also show initial findings from my research.

The examples I will use are based on the CEFR as applied to Irish, but some of the methods of analysis could apply to a wide range of languages. The Irish-language interpretation of the CEFR is chiefly realised by Teastas Eorpach na Gaeilge (European Certificate of Irish), or TEG. The sample materials provided by TEG for each CEFR level and the relevant syllabi will be used as a baseline from which the additional educational materials will be related to the CEFR.

Research of this type requires the consideration of multiple language features, from syntax to lexicon, and from grammar to discourse; a fact which is corroborated in Council of Europe (2009). The Manual goes on to state that it is not a blueprint, but

that it aims to encourage reflection and good practice when relating to the CEFR. The Manual also provides observations from its application in a pilot scheme; "[...] several users who piloted the preliminary edition commented that going through the procedures in the Manual was a good way to critically review and evaluate the content and the statistical characteristics of an examination — and that the outcome of this process was as important as the claim to linkage." (Council of Europe, 13:2009).

TEG's syllabi for levels A1 and A2 state that following are some of the language features that should be included in educational materials at that level. These stated language features will serve as baselines, and will be added to appropriate features identified in other literature.

The following features are described as being some of those that should be included at A1 and A2.

- An awareness of relevant grammatical terminology
- Syntactic differences and similarities between Irish and English and other languages
- Emphatic markers in Irish, as opposed to those in English
- Various plural endings for nouns
- Various endings for verbs in the future tense
- Using the imperative mood (with children)
- Examples of differences between dialects
- Phonetic differences between vowels and accented vowels
- Learners will understand single words and very simple sentences when discussing everyday life
- They will be able to communicate information about themselves, about the place they live, about their work, and things they do daily
- Simple and recognisable phrases
- And so on...

The publications discussed above give a brief introduction to that which has informed the selection and application of the following methods for relating educational materials to CEFR levels.

2 Methods chosen from research *on* educational materials

Römer (2006) provides several methods for pedagogical corpus-analysis. Comparison of semantically similar verbs, such as 'talk' and 'speak' or 'listen' and 'hear', can tell us if educational materials are giving a balanced view of subtle differences in the language and balanced evidence for learners to proceed in an informed manner. Römer (2006) also suggests comparing "problematic lexical-grammatical items" as they appear in both educational materials with how they are used by the language community; examples given for English include modals, tenses, connectors, verb-noun collocations, irregular verbs, future time expressions, linking adverbials, if-clauses, and the present perfect. While some of these language categories may be more problematic in one language than another, language experts should be able to select the categories their language's learners find most problematic. Römer (2006) reports that for each of the items investigated there was a mis-match between naturally-occurring English and English in educational materials. At this point, identifying which CEFR levels intersect with the features above seems the best way forward. Whether this is true or not will be examined in this research.

Hsu (2009) provides a corpus-analysis of general-English textbooks used in universities in Taiwan. In this research, vocabulary size and levels are analysed and compared with the BNC and Coxhead's Academic Word List (Coxhead: 2000). Hsu (2009) therefore aims to say whether learners who have completed Taiwan's language proficiency tests are properly equipped for the jobs market and naturally-occurring English. A key methodology used in Hsu (2009) is lemmatization. Lemmatization of the research corpus would certainly be of huge benefit to any morphologically complex language. Murakami (2009) also measures levels of vocabulary with Coxhead's Academic Word List, but also focuses on a comparison of vocabulary levels between educational materials used in different Asian countries. The 67 linguistic features investigated in Biber (1995). Biber (1995) collected these 67 features from research on English-language use, and in my research I will begin by using the areas of grammar TEG aligns with CEFR levels to relate educational materials with CEFR levels. Lists of vocabulary, and keywords, in materials can help researchers identify discourse markers or conduct automatic and semi-automatic comparisons between a reference list and target list. For a study of Irish materials, or any morphologically complex language, lemmatization of the data would ideally precede this stage of the research. A number of inflectional changes can be made to nouns in Irish depending on tense, case, number, and gender. See some examples below.

- shopping = siopadóireacht
- the shopping = an **t**siopadóireacht
- cost of the shopping = costas na siopadóireach**a**t
- your shopping = do s**h**iopadóireacht
- you were a shopper = ba s**h**iopadóir thú

In this talk I will focus on the pilot study results in

using known research methods and language features that have been specified in the syllabi for TEG's levels A1 and A2. I will report on how and why the language features above were, or were not, useful when relating educational materials to the CEFR.

3 Initial results

Sample lessons provided by TEG at A1 level include the images below and encourage the teacher to introduce the word "siopadóireacht" (shopping) as an activity, "siopa" (shop) as a place of work, and a lesson called "liosta siopadóireachta" (shopping list).

Shopping as an activity - A shop as a place of work - Write a shopping list

As we can see, semantically meaningful words such as 'shop' and 'shopping' are present in the TEG material, and can be used to mark certain types of discourse. It is clear, however, that a surface level comparison like this would not be sufficient and concordances should also be checked for sentence length and simplicity, among other features for level A1.

The following examples have been taken from the *Séideán Sí*, a publication for junior cycle in primary schools.

Ag siopadóireacht	le	mamaí
Shopping	with	mammy

This conforms to the following features specified in CEFR level A1:

- The household and family (with reference to 'mammy')
- Day-to-day activities (with reference to 'shopping')
- Shopping (with reference to 'shopping'!)
- Simple and short phrases or texts relating to everyday life

Séideán Sí can also be related at a discourse level to TEG's material for levels A1 and A2 in its lessons on '*Mo Thigh*' (My House); and with familiar phrases in another lesson called, '*is maith liom*' (I like). However, lessons teaching or introducing aspects of the future tense are not frequently found in *Séideán Sí* which has a much greater emphasis on the present tense or on short non-verbal statements that list some of the items in accompanying pictures. My talk will include more detailed results, and I will highlight considerations and implications arising from the relating process.

References

Biber, D. (1995) Dimensions of Register Variation: A cross-linguistic comparison. Cambridge University Press, 1995

Burton, G. (2009) Corpora and Coursebooks: destined to be strangers forever?

Council of Europe (2009) Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. http://www.coe.int/t/dg4/linguistic/source/manualrevision-proofread-final_en.pdf (downloaded 15 Dec, 2014)

Coxhead, A. (2000). A New Academic Word List. TESOL Quarterly, Vol. 34, No. 2 (Summer, 2000), pp. 213-238 http://www.jstor.org/stable/3587951

Meunier, F. (2002) The pedagogical value of native and learner corpora in EFL grammar teaching.

Teastas Eorpach na Gaeilge (n.d.), www.teg.ie (last accessed: Jan 12, 2014)

Ollscoil na hÉireann, Má Nuad (n.d.) Teastas Eorpach na Gaeilge, Syllabus A1. http://www.teg.ie/_fileupload/syllabi/A1_syll.pdf (last accessed: Jan 12, 2014)

An Gúm, (n.d.) Séideán Sí: Digital Resources. http://www.gaeilge.ie/about-foras-na-gaeilge/seidean-si/?lang=en (last accessed: 12 Jan, 2015)

Hypertextualizer:Quotation Extraction Software

Jiří Milička
Charles University, Prague
jiri@milicka.cz

Petr Zemánek
Charles University, Prague
petr.zemanek@ff.cuni.cz

1 Introduction

There are several strategies for automatic quotation extraction like searching for quotation marking phrases (for example c.f. Pareti et al. (2013), Pouliquen et al. (2007) and Fernandes et al. (2011)), or on the metadata processing (e.g. Shi et al. 2010). This approach does not take covert citations into account and as Kolak and Schilit (2008) from Google Research notice, it is hardly applicable to unstructured text corpora due to high variability in quotation marking styles.

The Google Books algorithm thus searches for all strings of words that repeat themselves which is "the most basic and reliable way to identify all quotations, although this will include some non-quotations" (Kolak et al. 2008).

Our Hypertextualizer goes a step further and introduces some tolerancy to the variability of these stings, namely the word order and some chosen percentage of word tokens. The algorithm is more complex but as implemented in our software, it is still applicable to huge corpora – it has been tested on a 420M word-token-long historical corpus of Arabic.

2 The Software

The software consists of two parts. The first one is designed to tokenize a raw text (adopting a pretokenized corpus is also possible), make indexation and search for similar word n-grams. As we intended to use the program to explore our Arabic corpora, the program is suitable not only for European languages, but for Arabic texts as well. As for the search algorithm, it is described in (Zemánek and Milička 2014). The user can specify desired tolerance rate and minimal length of quotations. The lower minimal length and higher tolerance, the more time the process takes and the more results it provides (however at some point, short repeating sequences tend to be common phrases and collocations rather than quotations).

The second part provides the opportunity to analyse the output data – to see quotations within a chosen subcorpus, sort those quotations according to certain parameters, view the corpus as a hypertext and export the links between texts into the *dot* format which is suitable for analysing and visualizing the networks by some external tools.

3 Studies Based on the Software

The first version of the software became functional in January 2014. In this early stage, it was not suitable for publishing; nevertheless, it enabled us to explore our aforementioned Arabic corpus.

The first study (Zemánek and Milička 2014a) focused on centrality of the hypertext network.

The second one (Zemánek and Milička 2014b) took advantage of hypertextual properties of the corpus in order to enhance the corpus search engine and to rank its results according to importance of the included texts.

These studies were also good opportunities to thoroughly test the algorithms as well as their practical implementation.

4 Acknowledgements

The research reflected in this article was supported by the GAČR (Czech Science Foundation), project no. 13-28220S.

References

Kolak, O. and Schilit, B. N. 2008. "Generating Links by Mining Quotations". In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*. New York.

Pareti, S., O'Keefe, T., Konstas, I., Curran, J. R. and Koprinska, I. 2013. "Automatically Detecting and Attributing Indirect Quotations". In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle.

Paul, W., Fernandes, D., Motta, E. and Milidiú, R. L. 2011. "Quotation Extraction for Portuguese". In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*. Cuiabá.

Pouliquen, B., Steinberger, R. and Best, C.. 2007. "Automatic Detection Of Quotations in Multilingual News". In *Proceedings of Recent Advances in Natural Language Processing 2007*. Borovets.

Zemánek, P. and Milička, J. 2014a. "Quotations, Relevance and Time Depth: Medieval Arabic Literature in Grids and Networks". In: *14th Conference of the European Chapter of the Association for Computational Linguistics*. 2014. Available online at http://aclweb.org/anthology//W/W14/W14-09.pdf

Zemánek, P. and Milička, J. 2014b. "Ranking Search Results for Arabic Diachronic Corpora. Google-like search engine for (non)linguists". In *Proceedings of CITALA 2014 (5th International Conference on Arabic Language Processing, Oujda)*. Available online at http://www.citala.org/papers/paper_29.pdf

Gender and e-recruitment: a comparative analysis between job adverts published for the German and Italian labour markets

Chiara Nardone
University of Bologna
chiara.nardone2@unibo.it

The e-recruitment phenomenon has changed the way companies address job seekers around the world, but, whereas numerous academic studies (Marschall 2002; Young et al. 2006) have focused on e-recruitment and its social, cultural and psychological effects, little is known about its linguistic features and about the related gender issues.

In Germany the subject "gender and language" has gained considerable interest among institutions and academia since 1978; as a matter of fact, the use of gender-fair strategies has gradually increased and the diffusion of generic masculine forms has diminished in the last 30 years. In Italy the debate around "gender and language" has received scarce attention both in the academic research and by institutions; indeed, generic masculine forms are still extremely common and accepted.

Even though numerous psycholinguistic studies (Gygax at al. 2008; Irmen 2007) have shown that using generic masculine forms for role names reveals a general male bias in the readers' and listeners' understanding, very few studies have analysed the way men and women are addressed in job adverts and which kind of consequences these forms of addressing have on labour markets.

The main purpose of this contribution is to investigate gender-biased forms and gender-fair alternatives used in job adverts published by German and Italian companies on their websites for the German and Italian labour markets.

The initial hypothesis is that gender-fair language is used more often in German job adverts than in Italian ones.

In order to test this hypothesis, a sample of job adverts has been collected from the career section of the websites of some German and Italian companies. Two comparable corpora have been built: one in German and one in Italian. Both corpora are composed by 260 job adverts published by 65 companies. Notwithstanding the same amount of job adverts, the two corpora have a different number of types and tokens, therefore the frequency of results has been normalized to a common base of 100,000 words.

The analysis on the corpora has been corpus-based rather than corpus-driven (Tognini-Bonelli 2001), in that the words chosen for examination were decided while reading and collecting job adverts. The recommendations contained in the guidelines on gender-fair language written by Robustelli (2012) for Italian and by Braun (2000) for German have also been followed.

The analysis on the frequency of masculine generic forms and on gender-fair alternatives has been carried out with the support of the corpus analysis toolkit *AntConc*. The results obtained represent the starting point for evaluating the cultural and linguistic elements that influence the way German and Italian companies communicate with job applicants in Germany and in Italy.

The analysis on this sample of job adverts shows that generic masculine forms are extremely common both in German and in Italian. The gender-fair strategies recommended by the guidelines are scarcely used: slash formulations, double formulations and gender-neutral words occur just in few job adverts both in German and in Italian.

However, in German job adverts "m/w" or "w/m" is often added to the generic masculine nouns in order to specify that job adverts address both women and men.

Furthermore, in German job adverts readers are addressed directly with the formal pronoun *Sie*: This strategy is generally recommended by guidelines for the use of gender-fair language and is very common in the analysed texts.

These results indicate that the initial hypothesis is only partially confirmed. Gender-fair strategies do appear more often in German job adverts, especially concerning the use of the pronoun *Sie,* but, at the same time, generic masculine forms still remain the most common alternative both in German and in Italian.

These initial findings can be connected with both linguistic and cultural reasons. On the one hand the different degree of attention to the "gender and language" debate given by Germany and Italy could explain why gender-fair forms are more used in German job adverts than in the Italian ones. On the other hand, the masculinity of both countries – according to Hofstede's cultural dimensions – may imply why generic masculine forms are still extremely common both in the Italian and in the German job adverts, even though there is scientific evidence that the use of these forms biases gender representations in a discriminatory way to women (Gygax et al 2008).

References

Braun, F 2000. *Leitfaden zur geschlechtergerechten Formulierung. Mehr Frauen in die Sprache.* Kiel: Ministerium für Justiz, Frauen, Jugend und Familie des Landes Schleswig-Holstein.

Gygax, P., Gabriel, U., Sarrasin, O., Oakhill, J. and Garnham, A. 2008. "Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men". *Language and Cognitive Processes*. 23(3), 464-485.

Hofstede, G., Hofstede, G.J. and Minkov, M. 2010. *Cultures and organizations: software of the mind: intercultural cooperation and its importance for survival*. New York: McGraw-Hill.

Irmen, L. 2007. "What's in a (role) name? Formal and conceptual aspects of comprehending personal nouns". *Journal of Psycholinguistic Research*. 36(6), 431-456.

Marschall, D. 2002. "Ideological discourses in the making of Internet career sites". *Journal of Computer-Mediated Communication.* 7.4.

Robustelli, C. 2012. *Linee guida per l'uso del genere nel linguaggio amministrativo.* Firenze: Comune di Firenze.

Tognini-Bonelli, E. 2001. *Corpus linguistics at work*. Amsterdam: John Benjamins Publishing.

Young, J. and Foot, K. 2005. "Corporate E-Cruiting: The Construction of Work in Fortune 500 Recruiting Web Sites". *Journal of Computer-Mediated Communication.* 11.1, 44-71.

Media reverberations on the 'Red Line': Syria, Metaphor and Narrative in the news "extended abstract"

Ben O'Loughlin
Royal Holloway, University of London
Ben.OLoughlin@rhul.ac.uk

Federica Ferrari
University of Bologna
federica.ferrari10@unibo.it

1 Introduction and theoretical context

Did Obama's 'red line' metaphor nearly trigger a military intervention in Syria in the summer of 2013? What work does the 'red line' metaphor do in shaping understandings and conduct in international affairs? The term is used by political leaders to express likely behavioural consequences to international rivals and allies and to domestic publics. What difference in diplomatic practice does it make to speak of a line, and a red one? How do such metaphors trigger or sustain narratives, and how do narratives lead to such metaphors? Last but not least, how was the 'red line' adjusted to avoid an international conflict and, as a result, to what extent has the 'red line' affected the leader's image and credibility?

The notion of the conceptual metaphor as developed by Lakoff and other researchers working within a cognitive approach to language and thought (Kovecses 2002; Lakoff 1993; Lakoff and Johnson 2003 [1980]; Steen 1999), operationally interacts with other analytical tools at a lexical, structural and narrative level - frames, discourse worlds (Chilton, 2004), and narratives. Interesting insight will emerge from a CADS approach (Corpus Assisted Discourse Studies, Stubbs 2001, Partington, 2008).

2 Panorama under analysis

In the context of conflict in Syria we examine the trajectory and remediation (Bolter and Grusin, 2000) of the red line metaphor and how actors use it to accomplish their objectives - to be seen to acknowledge, affirm, support, challenge, or subvert Obama's strategic narrative. We start from Obama's official declarations:

> We have been very clear to the Assad regime, but also to other players on the ground, that a red line for us is we start seeing a whole bunch of chemical weapons moving around or being utilized. That would change my calculus. That would change my equation. (Obama, 20 August, 2012)

> Let me unpack the question. First of all, I

> didn't set a red line; the world set a red line. The world set a red line when governments representing 98 percent of the world's population said the use of chemical weapons are abhorrent and passed a treaty forbidding their use even when countries are engaged in war.
>
> Congress set a red line when it ratified that treaty. Congress set a red line when it indicated that -- in a piece of legislation titled the Syria Accountability Act -- that some of the horrendous things that are happening on the ground there need to be answered for. (Obama, 04, September, 2013)

We take as an empirical nexus the September 2013 debate at the UN Security Council at which Samantha Power and her international peers discussed the consequences of military intervention in Syria, and consider the following series of official voices in correspondence with crucial delivery: Samantha Power's remarks on Syria in the UNSC debate (5 September, 2013); John Kerry's press conference remarks (13 September, 2013), and with Sergey Lavrov (12, September, 2013); Samantha Power's remarks "At a steakout [sic] on Syria" (16 September, 2013); Ban Ki-moon's remarks "on the report of the United Nations Missions to Investigate Allegations of the Use of Chemical Weapons" (16 September, 2013); Sergey Lavrov's speech at the UN general Assembly (27 September, 2013); John Kerry's remarks at UN Security Council (27, September, 2013); William Hague's and Asselborn's subsequent explanations (27 September, 2013).

We trace responses in international media (newspaper and television) with a particular focus on US vs UK newspapers' coverage in the period under observation.

3 Materials and methods

On the basis of the discoursal plethora of official voices considered, we select two US and two UK newspapers, belonging to different orientations and genre balanced (New York Times and Washington Times – US side; The Guardian and The Daily Mail – UK side). Corpus design criteria (Atkins & Clear 1992) are fundamental to define the materials under analysis, together with retrievability constraints. More specifically, the Corpus "Syria News 1309" is composed of 4 small mini corpora, each corresponding to the news coverage by The New York Times , the Washington Times, the Guardian and the Daily Mail over September 2013 (cf. Sibol2013 Corpus). All the news are considered in the period between 4 and 28 September 2013 to observe the connotational, argumentational and rhetorical behaviour of "red line" across the news and along the period under consideration. Also under analysis is how the metaphor has been moved and changed according to context, time and political actions and to what extent it has affected Obama's image as an international leader.

4 Analysis' results and discussion

Corpus analysis of "red line*" (searching for "red", sorted by 1R, with focus on "line*") gave rise to the emergence of 13 occurrences in the Daily Mail, 41 occurrences in the Guardian, 108 occurrences in the Washington Times, 88 in the New York Times.
Concordance analysis allows us to investigate the complexity of the political case at issue and observe potential backlashes on the leader's image. See for instance the following example: "*Obama may have fallen victim to his "red-line" bravado, but he has drawn Russia into closer involvement*" (Concordance 303, The Guardian, 18 September, 2013). Also, ideological positioning emerges from debate reporting in accordance with the news political orientation and the contextual articulation of the case in point. See for instance the following example: "*Mr. Obama has gotten by until now with redefining reality as what he says it is. Red line? What red line? Now he wants to similarly redefine war…*" (Concordance 31, The Washington Times, 10 September, 2013).

The significance of our argument is to open up reflection on the function of metaphor and narrative in steering sense-making in diplomatic practice and to highlight its pragmatic force and dynamics along various degrees of genre variation (official vs. news voices) within the complexity of discourse as an ever changing interactive mutual vocal practice.

5 References

Atkins, S. & J. Clear. 1992. Corpus design criteria. *Literary and Linguistic Computing, 7*(1), 1-16.

Bolter, J. D. & Grusin, R. A. 2000. *Remediation: Understanding new media.* Boston, MA: MIT Press.Chilton, P.A. 2004. *Analysing political discourse*. London and New York: Routledge.

Kovecses, Z. 2002. *Metaphor: A Practical Introduction*. Oxford: Oxford University Press.

Lakoff, G. 1993. The Contemporary Theory of Metaphor. In Ortony, Andrew (ed.) *Metaphor and Thought* (2nd edition). Cambridge: Cambridge University Press, 202-251.

Lakoff, G. & M. Johnson 2003 [1980]. *Metaphors we Live By*. Chicago: University of Chicago Press.

Partington A. 2008. The armchair and the machine: Corpus-Assisted Discourse Studies, in C. Taylor Torsello, K. Ackerley and E. Castello (eds) *Corpora for University Language Teachers*, Bern: Peter Lang, 189-213.

Stubbs, M. (2001). *Words and phrases*. Oxford: Blackwell.

Exploring the language of twins: a corpus-driven pilot study

Carlos Ordoñana
University of Murcia.
carlos.ordonana@um.es

Pascual Pérez-Paredes
University of Murcia.
Pascualf@um.es

The acquisition of a second language (SLA) is affected by several factors at once. While some of them are purely dependent on the environment in which the learner acquires the language, every person possesses a number of individual factors, such as motivation, language aptitude, language transfer and cognitive maturity (Paradis 2011). In a similar way, the attitude of the learner also affects the way the language is acquired. As many of these individual factors have a high genetic component that varies from one person to another, determining the influence of genes on the process of language acquisition could open a wide range of possibilities in the field of SLA.

In order to explore the genetic component of the individual's L1 language acquisition, scientists make use of the concept of heritability (Plomin 2012), which could be defined as the proportion of the variance in a population that is related with the genetic heritage, assuming that all that variance between subjects within a group is due to differences in the genetic and environmental factors that affect each individual. The most efficient way of studying heritability is by using twin couples as subjects, as they share genetic material, as well as early environment. Comparing monozygotic twins (which share 100% of their segregating genes) with dizygotic twins (which share on average 50% of their genetic makeup) could facilitate the identification of the environmental and genetic factors.

The use of corpus linguistics methods may be instrumental in determining the degree of similarity in the production of L2 texts. In order to test the feasibility of such methodology, we designed a pilot study. We collected a small corpus of texts (500 words limit each) written in English in which the informants had to explain their agreement with the statement "*Most university degrees are theoretical and do not prepare students for the real world*", one of the suggested essay topics in the *The International Corpus of Learner English* (ICLE). The subjects were 12 couples of monozygotic (7 pairs) and dizygotic twins (5 pairs); all of them were university students whose age ranged from 19 to 24 years. With only a couple of exceptions who studied

French as FL, all of them had acquired the English language in an instructed environment. Two independent language instructors evaluated each text following the Common European Framework of Reference for Languages (CEFR). The average level of proficiency was mostly intermediate (B1-B2). We compared the evaluations between the members of each pair, and combined with the use of tools such as *Wordsmith* (Scott 2008) *or Lextutor,* we aimed to explore the main features of the language used by each pair.

Due to the limited quantity of subjects we had at our disposal, definite conclusions cannot be drawn from the results obtained. However, the exploration of the methodology used may allow us to improve the future attempts of further corpus-driven research in the field of heritability regarding SLA.

References

Paradis, J. 2011. Individual differences in child English second language acquisition: Comparing child-internal and child-external factors. *Linguistic Approaches to Bilingualism, 1*,3, 213-237.

Plomin, R., DeFries, J.C., Knopik, V.S., Neiderhiser, J. M. 2012. *Behavioral Genetics* (6 ed.). New York: Worth Publishers.

Scott, M. 2008. Developing WordSmith. *IJES, International Journal of English Studies, 8*,1, 95-106.

Mono-collocates: How fixed Multi-Word Units (MWUs) with *of* or *to* indicate diversity of use in different corpora.

Michael Pace-Sigge
University of Eastern Finland
michael.pace-sigge@uef.fi

1 Introduction:

There is evidence that particular words in the English language do not freely collocate; similarly not all words of a single word class fit in the same colligational structure. Thus, John Sinclair and others eg. Stubbs (1996) pointed out that commonly found structures often rely on a specific node word. Corpus-based investigations, provide evidence for this, as the following example from the Collins Cobuild Grammar shows:

> There are a few adjectives which are always or almost always used in form of a noun and are never or rarely used as the complement of a link verb. These adjectives are called attributive adjectives. Examples are 'atomic' and 'outdoor'. You can talk about 'an atomic explosion', but you do not say, 'The explosion was atomic'. You can talk about 'outdoor pursuits', but you do not say 'Their pursuits are outdoor'.
>
> (Sinclair, 1990: 80)

Francis, similarly, points out that "where the introductory, or non-referring pronoun it is the Object of a verb, and is followed by an adjective or noun group (…) the structure occurs with an extremely restricted range of verbs, of which *find* and *make* are by far the most frequent, accounting for over 98% of all the citations of the structure in the corpus." (Francis, 1993: 140f.). This paper is concerned with fixed bigrams using the items of and to. The focus is on those items (words) which are close collocates for of and to and which appear to have a total word count that is not much higher than the total number of occurrences of these words as bigrams. An alternative approach is to look at items which have of or to as a near-collocate that outnumbers the next most frequent collocate by at least the factor of 100:1.

2 The Investigation

In a detailed investigation of the uses of the items *of* and *to* (Pace-Sigge, forthcoming) it has become apparent that there are, in fact, a number of node words that are close collocates with these items. It can be said that such two-word units are, being found far more often than other possible collocates,

mono-collocational. To name an item *mono-collocate,* it should appear bound to either *to* or *of* eight (or more) out of ten times. Alternatively, a single item can be deemed mono-collocational if other collocates are of low relative frequencies.

In this paper, *of* and *to* usage has been investigated in six corpora: two are casual spoken British English (spontaneous spoken), a further two look at public speeches (prepared-spoken), and a final two look at British fiction of the 19th and 20th century.

Such mono-collocates appear to create MWUs which that are essential building blocks of communication. There are a number of items, for example *used, able* (with *to*) or *kind, sort* (with *of*) which show strong tendencies to be mono-collates overall. The strength in their bonds can either be genre- or corpus- specific; on the other hand, a number of word units appear regardless of genre.

Pragmatic and stylistic needs expressed by such usage patterns can be explained with reference to *priming* processes - see Hoey (2005) - whereby repeat exposure sets a template for sets of words to appear in a fixed construction in the majority of cases. With reference to Hoey's theory, bigrams can be disambiguated by the item that appears directly next to the node-word: hence *the sort* (classification), *sort out* (separation) and *sort of* (discourse particle). These three examples can be traced to the same semantic root, yet are employed in separate ways; the one furthest removed from the meaning of the node happens to be the one predominantly found. Similarly, it could be argued that, where such texts are for a specific audience (as in public speeches), listeners are primed to expect particular sets of words.

3 Conclusion

The existence of lexical items that are occurring in multi-word format can be seen as salient within the English language. This is an issue that is potentially relevant to the teaching of the language, in a way similar to the teaching of phrasal verbs.

Mono-collocates can, in particular, express pragmatic needs (*seems to* and *of course* in the public speech data) or, in general, function to express narrative time-frames with phrasal verbs (*used to, want to, going to*).

Mono-collocations are also found in texts to employ the use of vagueness markers on the speaker's (or author's) part, when a more detailed description is either not needed or wanted (*lot of, kind of, sort of*). The data investigated shows that such bigrams can, on occasion, claim 100 per cent (or close to 100 per cent) of all uses of a specific item. Even where there is no 100 per cent lock-in with a single collocate, the colligational structure (for example, being of the same word-class like, pronoun) appears to cement *of* or *to* into a fairly fixed MWUs for a number of key node words.

References

Francis, G. 1993. A Corpus-Driven Approach to Grammar Principles, Methods and Examples. In: Baker, Mona, Francis, Gill, and Tognini-Bonelli, Elena, eds. Text and Technology : In Honour of John Sinclair. Amsterdam, NLD: John Benjamins Publishing Company, 1993.

Hoey, M. 2005. *Lexical Priming*. London: Routledge.

Pace-Sigge, M. (forthcoming) *The Function and Use of TO and OF in Multi-Word Units.* Hounslow: Palgrave Macmillan.

Sinclair, J. (Editor-in-Chief) et al. 1990. *Collins Cobuild Grammar.* London: Collins.

Sinclair, J. [1992] 2004. *Trust the text. Language, corpus and discourse*. London: Routledge.

Stubbs, M. (1996). *Text and Corpus Analysis.* Computer-Assisted Analysis of Language and Culture. Oxford: Basil Blackwell.

Streamlining corpus-linguistics in Higher and adult education: the TELL-OP strategic partnership

Pascual Pérez-Paredes
Universidad de Murcia
pascualf@um.es

1 Introduction

The European Space of Higher Education (HE) and the CEFRL demand new teaching and learning methodologies that promote more active participation of the language learner. Learning is increasingly turning into a learner-centered process where the needs of the learners are catered for, and the teacher, if any, acts as a guide or facilitator while students become proactive subjects (Pérez-Paredes & Sánchez Tornel, 2009). Besides, learning can happen anywhere, anytime. This "new" mobile learning is increasingly popular, and a cost-effective way to meet the needs of masses of people (Kinshuk, Huang & Ronghuai, 2015.).

Given these new possibilities, wouldn't it make more sense if adult learners could personalize their learning and use their own language output (using their text and their own voices) to further acquire language skills? How can learners take advantage of their own mobile devices to input their own language and gain further communicative competence? Can we personalize language learning by taking advantage of Natural language processing (NLP) services and technologies already available? How can adult learners use their critical thinking, analysis & awareness skills (Aguado-Jiménez, Pérez-Paredes & Sánchez, 2012) to improve their communicative competence (Pérez-Paredes, 2010) across different CEFR levels by using these open educational resources (OER) tools?

2 The TELL-OP rationale

TELL-OP is a transnational Strategic Partnership that involves at this point five HE organizations from different countries and which seeks to produce innovative outputs in the fields of both HE & adult foreign language learning by addressing the new agenda on HE and lifelong learning & the needs for labour market skills.

Clearly, there is a demand for mobile learning and those in HE & adult language learning education have the obligation to provide the opportunities for the use of OERs that are adapted to mobile ubiquitous learning based on evidence-based good practices across levels (A2 and B2 in the case of TELL-OP) and languages. We aim at maximizing the role of learner language by promoting good practices in using these OERs in personalized language leaning contexts and thus contribute to the modernization of the HE systems in the EU and a more widespread use of innovative OERs and learning designs (Conole, 2013) that include not only English but also other EU languages that can serve as the basis for a more widespread use of these ICTs.

TELL-OP is a Strategic Partnership that seeks to promote the take-up of innovative practices in European language learning (Data Driven Learning, DDL) (Boulton & Pérez-Paredes, 2014) by supporting personalised learning approaches that rely on the use of ICT and OER by bringing together the knowledge and expertise of European stakeholders in the fields of language education, corpus and applied linguistics, e-learning and knowledge engineering in order to promote cooperation and contribute to unleash the potential behind already available web 2.0 services to promote the personalized e-learning of languages in the contexts of higher and adult education, in particular, through mobile devices.

Instead of producing these OER resources, the TELL-OP consortium is interested in finding existing NLP OER that can suit the needs of language learner across different European languages (English, German and Spanish) and learning scenarios (Adult and HE education) and streamline these services by carrying out an exchange of good practices and evidence-based research that is focused on learners' needs and not so much on context-free academic endeavours.

3 Aims

The objectives of TELL-OP are the following: (1) to promote the use of learner language information in the context of higher and adult education in Europe by offering concrete models of use that can be taken up by our target groups; (2) to survey and document the most relevant OE resources and services for language processing (text and voice) in the context of higher and adult education in Europe. That means, analyzing needs in the EHEA and primarily in the stakeholders' countries and establish the starting point according to needs; (3) to raise awareness on the usefulness of using learner language input for the learning and teaching of languages in Europe in the 2 scenarios outlined in this proposal: formal HE and informal adult language education; (4) to promote a cluster group of EU experts and professionals who can bring together their different views and expertise in the fields of e-language learning, language education, corpus linguistics and knowledge engineering.; (5) to foster the usage of the OERs and ICT-mediated

language processing methods for the creation of the language information suitable for pedagogic purposes in English, Spanish and German.

Acknowledgements

Transforming European Learner Language into Learning Opportunities 2014-1-ES01-KA203-004782, a KA200 Higher Education Strategic Partnership, funded by the OAPEE and the EU.

References

Aguado-Jiménez, P., Pérez-Paredes, P., & Sánchez, P. (2012). Exploring the use of multidimensional analysis of learner language to promote register awareness. *System*, 40(1), 90-103.

Boulton, A., & Pérez-Paredes, P. (2014). Researching uses of corpora for language teaching and learning Editorial Researching uses of corpora for language teaching and learning. *ReCALL*, 26, 121-127.

Conole, G. 2013. *Designing for Learning in an Open World*. Explorations in the Learning Sciences, Instructional Systems and Performance Technologies, Vol. 4. Springer.

Kinshuk, Huang, Ronghuai (Eds.). 2015. *Ubiquitous Learning Environments and Technologies*. Lecture Notes in Educational Technology. Springer.

Pérez-Paredes, P. 2010. Corpus Linguistics and Language Education in Perspective: Appropriation and the Possibilities Scenario. In T. Harris & M. Moreno Jaén (Eds.), *Corpus Linguistics in Language Teaching* (pp. 53-73). Peter Lang.

Pérez-Paredes, P., & Sánchez Tornel, M. (2009). Understanding e-skills in the Foreign Language Teaching context: Skills, strategies and computer expertise. . In R. Marriott & P. Torres (Eds.), *Handbook of Research on E-Learning Methodologies for Language Acquisition* (pp. 1-22). IGI Global.

Conditionals and verb-forms in nineteenth-century life-sciences texts

Luis Miguel Puente Castelo
Universidade da Coruña
luis.pcastelo@udc.es

Begoña Crespo García
Universidade da Coruña
bcrespo@udc.es

Conditionals are a particularly valuable resource in scientific register (Carter-Thomas & Rowley-Jolivet 2008: 91) as they can fulfil an important number of different functions, both to highlight relations between the premises of the discourse and to establish cooperative links between the audience and the author, among others. This versatility is in part a result of the very important degree of variability of conditionals, which occurs at several levels: Conditionals may be introduced in very different contexts, and their constituents can appear in different positions inside the structure. Moreover, they can be introduced using a large number of different subordinators, such as if, unless or as long as, as well as the inversion of particular operators, such as had or were, as shown in examples such as "Had he not seen the car coming, he would have been killed" (Biezma 2011: 555).

However, perhaps the most notorious example of the variability of conditional structures and of its relation with their meaning is not any of these, but the choice of combinations of tenses in both constituents of the conditional structure. In fact, most traditional and EFL grammars considered the different combinations of tenses as the main criterion to distinguish among several types of conditionals, which led them to promote a model presenting three types of conditionals encoding three degrees of hypotheticality: first-type (present simple + will), second-type (past simple + would), and third-type (past perfect + would have) conditionals.

This model has been thoroughly criticised (Hwang 1979, Maule 1988, Fulcher 1991, Ferguson 2001, Jones & Waller 2010), on the basis that it does not reflect the real use of conditional structures (especially in scientific writing) and that it sacrifices the variability of conditionals for the sake of easing the task of learning the structure for EFL students. Moreover, several corpus-based studies on conditionals in scientific writing have found that the three-type model only accounts for a very small portion of conditionals. For instance, the three canonical types combined account for only 14.7% of the occurrences of conditionals in Carter-Thomas & Rowley-Jolivet's corpus (2008: 195) and for just 18% in Ferguson's (2001: 70).

These analyses, however, have focused on present-day scientific writing, and it is not known whether the situation was different in previous stages of its development. Thus, the aim of this poster is to present the preliminary results of an analysis on the use of the different combinations of verb tenses and modals in nineteenth-century academic writing, using life-sciences texts as an example.

This research has been carried out using the nineteenth-century section of CELiST, one of the subcorpora of the Coruña Corpus of Scientific Writing (Crespo & Moskowich 2010; Moskowich 2011). The Coruña Corpus is a corpus containing samples of scientific texts from 1700 to 1900 which consists of several twin subcorpora, dealing with different disciplines and presenting the same design and principles of compilation. For this study, the twenty nineteenth-century samples of CELiST, totalling c. 200,000 words, have been used. This corpus has been searched for conditional particles (Quirk et al. 1985) with the Coruña Corpus Tool (Moskowich & Parapar 2008), a concordancer specifically designed to work with the Coruña Corpus, and the results have been then manually disambiguated in order to eliminate all non-conditional uses of the particles from the list of occurrences.

Once the disambiguation has been completed, each occurrence has been classified according to the verb forms in both constituents, and the results have then been analysed, looking both at the general use of verb tense combinations and at the possible correlation of this use with some variables, such as the type of conditional being used, the sex and geographical origin of the authors, the genre of the texts, or the year of publication, in order to find possible factors explaining the distribution of the uses.

Acknowledgements

The research here reported on has been funded by the Consellería de Educación e Ordenación Universitaria (I2C plan, reference number Pre/2011/096, co-funded 80% by the European Social Fund) and the Ministerio de Economía y Competitividad (MINECO), grant number FFI2013-42215-P. These grants are hereby gratefully acknowledged.

References

Biezma, María. 2011. "Conditional inversion and givenness". *Proceedings of SALT* 21: 552-571.

Carter-Thomas, Shirley & Elizabeth Rowley-Jolivet. 2008. "If-conditionals in medical discourse: from theory to disciplinary practice". *Journal of English for Academic Purposes* 7: 191-205.

Crespo, Begoña & Isabel Moskowich. 2010. "CETA in the Context of the Coruña Corpus". *Literary and Linguistic Computing* 25/2: 153-164.

Ferguson, Gibson. 2001. "If you pop over there: a corpus-based study of conditionals in medical discourse". *English for Specific Purposes* 20: 61-82.

Fulcher, Glenn. 1991. "Conditionals revisited". *ELT Journal* 45: 164-168.

Hwang, Myong Ok. 1979. *A semantic and syntactic analysis of if-conditionals*. Unpublished MA thesis. University of California Los Angeles.

Jones, Christian & Daniel Waller. 2010. "If only it were true: the problem with the four conditionals". *ELT Journal*. doi: 10.1093/elt/ccp101

Maule, David. 1988. ""Sorry, but if he comes, I go": Teaching conditionals". *ELT Journal* 42: 117-123.

Moskowich, Isabel. 2011. ""The golden rule of divine philosophy" exemplified in the Coruña Corpus of English Scientific Writing". *Revista de Lenguas para Fines Específicos* 17: 167-197.

Moskowich, Isabel & Javier Parapar. 2008. "Writing science, compiling science: The Coruña Corpus of English Scientific Writing". In María Jesús Lorenzo Modia (ed.) *Proceedings from the 31st AEDEAN Conference*. 531-544. A Coruña: Universidade da Coruña.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Studying the framing of the Muslim veil in Spanish editorials

Ricardo-María Jiménez
Universitat Internacional de Catalunya
rmjimenez@uic.es

1 Introduction

Throughout 2010, the problem of the Muslim veil was current in Spain. In April 2010, Najwa Malha, a teenage girl who studied at Camilo José Cela Secondary School (Pozuelo, Madrid, Spain), attended her school wearing a Muslim veil. The head of the school had forbidden the wearing of this kind of veil because the school regulations do not allow students to wear head coverings.

2 Theoretical background, methodology and corpus

This paper undertakes a linguistic and framing analysis of the social debate concerning the regulation of the wearing of the Muslim veil in public —in other words, the use of the *burqa* and *niqab*, both of which cover the wearer's face— and, by extension, the presence of religious symbols in the public sphere. The social debate is studied in a corpus of editorials from four major national newspapers in Spain belonging to clashing ideologies: *ABC* (right-wing), *El País* (left-wing), *La Vanguardia* (center) and *El Periódico* (left-wing).

I study editorials –the voice of the newspaper (Morley 2004)– because "they play a crucial role in the formation and change of public opinion, in setting the political agenda, and in influencing social debate" (Dijk 1996).

I used corpus methodology to uncover the frames, complementing my qualitative reading by bringing them out into the open. Sketch Engine tools were used to process the texts and identify key words and terms that might shed light on the ideological framing of the issues.

The subcorpus of editorials is composed of 154,173 tokens. The whole corpus of editorials is 2,692,837 tokens.

To select editorials related to the topic of the presence of religion in public spaces, a list of keywords was drawn up, with Sketch Engine tools, related to religious issues and the persons involved. The manually-selected items formed subcorpora of four newspaper editorials.

Sketch Engine allows the extraction of candidates for collocations (terms) in the subcorpus of editorials. However, candidates found by Sketch Engine are not properly-speaking collocations, but terms referring to people involved and relevant facts.

I analysed terms (Table 1) that are expressions or common syntagmas.

I think critical discourse analysis have often been used for their "impressionistic" or "manipulative" accounts of media text (Breeze 2011), so methods normally associated with corpus linguistics can be effectively used by critical discourse analysts (Baker and al 2008).

Techniques of frame analysis are used to identify and evaluate how issues involving the religious symbols worn by Muslims (Muslim veil) in Spain are discussed.

Since the early 1990s there has been a steady growth in the use of frame analysis in research about news and journalism, in an effort to offer insight into the forces that shape media interpretations of reality and their potential influence on audiences. The roots of framing as a theory are situated in the field of sociology, where the term has been in use since the mid-1950s (Bateson 1955). It is this sociological approach to framing that has underlain the study of news frames so far, with frames being examined as social constructs and outcomes of journalistic norms or organisational constraints, as well as sponsored by social and political actors. Goffman's (1974) prominent formative work has been pivotal in this field. Goffman defines frames as 'schemata of interpretation' that enable individuals to understand certain events and 'to locate, perceive, identify and label' occurrences. He calls these schemata primary frameworks because 'they turn what should be a meaningless aspect of a scene into something meaningful' (pp. 21–22).

Entman (1993) states frames are not only in texts, but also in the emissary, the receptor and in culture.

An important implication for media and communication research is that the coverage of events in the news media depends on the frameworks employed by journalists (Scheele, 2000).

The researcher has to examine the texts to detect possible framing elements, quantify their presence in the texts, and then explore the way in which these elements cluster together in order to assess the extent to which they form coherent frames containing all or most of the elements outlined by Entrant (1993).

Contreras (2004) analysed framing in journalistic discourse about the Catholic Church in international press in an extensive study that follows traditional journalistic analysis.

The paper seeks to demonstrate the contribution that corpus linguistic software can make in the frame analysis of editorials and how it can help address some of the methodological challenges in the study of frames. One of the first studies applying corpus linguistic software in framing was that of Touri and

Koteyko (2014).

3 Results and discussion

Term 'velo integral' (Muslim veil, *burqa* and *niqab*) appears as shown in Table 1.

Newspaper	Term 'velo integral'
El País	11
El Periódico	2
La Vanguardia	6
ABC	4

Table 1

In the editorial of **El País** the **issue** is a ban of the full veil in public places.

The full veil is **defined** as an article of clothing that discriminates against women –an aspect which is not abandoned in subsequent editorials– and as a symbol of religious expression. The first element of the definition remains in editorials; however, the second element definition changes: first, it is mentioned that is a symbol of religious expression and then later this is rejected. Perhaps this change is due to the difficulty of presenting a frame regarding an item of wear without a tradition in Spain which has caused a heated social debate.

The **evaluation** of the problem is a very thorough one. There is reference to electoral interests which lie behind the debate. In this editorial the problem is set out and there is discussion to determine the most appropriate solution. The integration of Muslims is only touched on in the editorial. Although it favours a ban on the full veil, the writer of the editorials clarifies many different solutions in the case of the veil or headscarf, which raise any more questions, perhaps for fear of provoking the fundamentalists, as stated in an editorial. The debate on the headscarf is a subordinate frame.

In the editorials of **El Periódico** the **issue** is that banning the full veil will backfire.

In El Periódico the *burqa* and the veil or headscarf are **defined** as religious symbols and that element of the definition remains in editorials.

Regarding **evaluation**, frame elements of both veils are included within the same group. Several editorials the ramifications of a headscarf ban are contemplated: such as how the Muslim community might be integrated into European society. On the other hand, as in the editorial of El País, although less frequently, there is mention of the fact electoral considerations may have pushed politicians to ban the veil. As for **solutions**, it is argued that there are reasons to ban it and reasons for not doing so. In any case the key is discussion to find the best solution and not cause adverse reactions among Muslims.

The **issue** in the editorials of **ABC** is that the veil should be banned.

In these editorials the full veil and the veil or headscarf is not **defined** as a religious symbol.

The evident problem in the frame elements in the editorials of ABC is to ban any kind of veil. They state that the *burqa* is a garment of cultural significance and then subsequently to deny it. As for the **evaluation,** there is frequent mention of the false liberalism of those who want to allow their use and fallacies made by those who defend the veil and then defend the ban Christian symbols in public spaces. It supports its argument to the rule of law, democratic principles and legality. In the first editorial of ABC one of the reasons to ban their use is the rejection of shelters, but that case was subsequently abandoned and not re-quote in the remaining publishers. The necessary integration of Muslims is mentioned. There is a reference in editorials of ABC to multiculturalism, which is the excuse used by those who want to allow their use. It highlighted in the editorials of ABC more solutions than those offered by El País and El Periódico.

The **solution** for ABC is clear: we must ban the headscarf, both the full veil as the Islamic headscarf.

The **issue** in the editorial of **La Vanguardia** is that the headscarf should be banned.

In the Editorials of La Vanguardia only **defined** in a text Muslim veil is equivalent to wearing a cross religious symbol, but not the frame element shown in the remaining texts.

On **evaluation**, it is emphasized the authorities should take action on the matter, and do not leave it in the hands of the Municipalities. It stresses the need for Muslims to integrate, as proposed by El Periódico, ABC and slightly El País. Also it goes to the ambiguous sense as a place of reference to know why or why not is to prohibit element also it is mentioned in other newspapers. It refers to the concept of multiculturalism, in this case paper with positive value and why we live in Europe in a multicultural society.

About **solutions**, it is proposed to be flexible with headscarf lets face uncovered and prohibit the wearing of full veil.

4 Conclusions

The paper reflects the usefulness of frame analysis to explore the media representation of controversial questions of this kind and the relations between discourse and ideology.

The *Journalism trenches* (López-Escobar 2008) of El País and ABC is clearly reflected in the topics related to Christian/Catholic, but the ideological polarization of the newspapers is quite blurred in the treatment of Muslim veil, that is, in a religion that is not Catholic. The polarization is quite blurred in La Vanguardia and El Periódico in matters of both the

Catholic and Muslim religions.

There is a correlation between the ideological line arguments newspapers and editorials.

Frame analysis needs complemented with a textual and linguistic analysis detailed discoursive. Therefore the analysis of the frames will delve into literalism, formulation and structure of the text.

References

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyzanowski, M., McEnery, T., and R. Wodak 2008. "A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press". *Discourse Society*, 19: 273-306.

Bateson, G. 1955. "A theory of play and fantasy". *AP: Psychiatric Research Reports*, 2: 39–51.

Breeze, R. 2011. "Critical discourse analysis and its critics". *Pragmatics*, 21/4: 493-525.

Breeze, R. 2013. "British media discourses on the wearing of religious symbols". In H. Van Belle, Gillaerts, P., Gorp, B. van, Mieroop, D. van de and K. Rutten (eds.) *Verbal and visual rhetoric in a media world*. Leiden: Leiden University Press, pp. 197-211.

Contreras, D. 2004. *La iglesia católica en la prensa: Periodismo, retórica y pragmática*. Pamplona: Eunsa.

Contreras, D. 2014. "The crucifix and the Court in Strasbourg: press reaction in Italy to a European court decisión". In I. Olza, Loureda, O. and M. Casado-Velarde (eds.) *Language Use in the Public Sphere: Methodological Perspectives and Empirical Applications*. Frankfurt am Main: Peter Lang: 327-350.

Dijk, T. van 1996. "Opinions and ideologies in editorials". *Paper for the 4th International Symposium of Critical Discourse Analysis, Language, Social Life and Critical Thought*. Athens, 14-16 December, 1995.

Entman, R. M. 1993. "Framing: Toward clarification of a fractured paradigm". *Journal of Communication*, 43(4): 51-58.

Goffman, E. 1974. *Frame analysis: An essay of the organization of experience.* New York, NY: Harper and Row.

López-Escobar, E. et alii 2008. "Election News Coverage in Spain: From Franco's Death to the Atocha Massacre". In Kaid, L. L., *The Handbook of Election News Coverage around the World.* New York: Routledge: 175-191.

Morley, J. 2004. "The Sting in the tail: Persuasion in English editorial discourse". In Partington, A., Morley, J. & Haarman, L. (eds.) *Corpora and Discourse*. Bern: Peter Lang: 233-252.

Olza, I. 2014. "Representations in the Spanish Press of the Political Debate about Wearing Full Islamic Veils in Public Spaces". In Olza, I., Ó. Loureda & M. Casado-Velarde (eds.) *Language Use in the Public Sphere: Methodological Perspectives and Empirical Applications*. Frankfurt am Main: Peter Lang: 521-547.

Scheufele, D. T. 2000. "Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication". *Mass Communication & Society*, 3: 297-316.

Touri, M. and N. Koteyko 2014. "Using corpus linguistic software in the extraction of news frames: towards a dynamic process of frame analysis in journalistic texts". *International Journal of Social Research Methodology*: 1-16.

Multi-functionality and syntactic position of discourse markers in political conversations: The case of 'you know', 'then' and 'so' in English and 'ya'nĩ' in Arabic

Saliha Ben Chikh
University of Paris III
Sorbonne Nouvelle
slh.benchikh@gmail.com

Each language is organized in accordance with its culture; it follows the social purposes of the interactants within that culture. The main concern in this paper is to reveal the extent to which discourse markers like *'you know'* and *'ya'nĩ'* (I mean/It means), *'then,* and *'so'* are pragmatically multifunctional in English and Arabic political conversations.

Using a corpus based approach, our study analyses material from political interviews aired on CNN and Aljazeera. After selecting and sorting the linguistic data with the aid of the aConCorde tool, our study involves three steps: a syntactic analysis identifying the positions of the markers, a semantic analysis distinguishing their uses and a pragmatic analysis defining their functions in political verbal interactions.

We assume that these expressions are inherently related to social criteria, context and syntactic position. The relationship between participants is also of paramount importance in our analysis of discourse markers. Our framework thus makes use of pragmatic theories about language (Erman 2001, Brinton 1996, Brown & Levinson 1987, Blakemore 2002, Dostie 2004, Leech 1980, and Kerbrat Orecchioni 1990, 1992, 1994).

Different scholars have variously emphasized the role and functions of discourse markers in texts and conversations in both French and English. However, this issue is not deeply treated in Arabic, mainly in the area of verbal exchanges. Speakers of Arabic use some pragmatic expressions massively in their daily conversations such as*:* ***ya'ni*** (that is/ I mean/ you mean/It means, well/so…), a'taqid (I believe/I think), aẓunnu (I suppose/ I think), wa lakin (but), al-muhim (important), bas (but), wallahi (indeed/ well)…etc.

Our findings indicate that *'you know'* and *'ya'nĩ'*, *'then,* and *'so'* can be used differently from one speech situation to another and from one position to another; they perform a range of interpersonal and interactional functions. Providing a variety of meanings, these pragmatic unites are thus strongly poly-functional and play an essential role in political conversations.

Strongly enough, these pragmatic units have undergone a pragmaticalization process, which gives them a set of inferential meanings; this process contributes to the acquisition of a variety of pragmatic functions in different contexts. The markers *'you know'* and *'ya'nĩ'* *'then'* and *'so'* have interpersonal and interactional purposes: they soften the force of an illocutionary act and derive relevant inferences of implicit meanings.

In both Arabic and English, speakers are found to use them for a variety of pragmatic contexts to perform auto-correction, reformulation, hedging and mitigating face-threatening acts, holding a turn, and request for implication and cooperation of interactants.

References

Blakemore, D. (2002). Relevance And Linguistic Meaning: The Semantics And Pragmatics Of Discourse Markers. Cambridge: Cambridge University Press.

Brinton, L, J. (1996). Pragmatic Markers In English: Grammaticalization And Discourse Functions. Herndon: Walter De Gruyter.

Brown, P., Levinson, S. C. (1987). Politeness- Some Universals In Language Usage. Cambridge: Cambridge University Press.

Dostie, G. (2004). Pragmaticalisation Et Marqueurs Discursifs: Analyse Sémantique Et Traitement Lexicographique. Bruxelles: Duculot.

Erman, B. (1987). Pragmatic Expressions In English, A Study Of « You Know », « You See » And « I Mean » In Face-To-Face Conversation, Doctoral Dissertation At The University Of Stockholm. Stockholm.

Erman, B. (2001). "Pragmatic Markers Revisited With A Focus On You Know In Adult And Adolescent Talk". Journal Of Pragmatics 33: 1337-1359. Elsevier Science B.V.

Kerbrat- Orecchioni, C. (1990). Les Interactions Verbales I. Paris: Armand Colin.

Kerbrat- Orecchioni, C. (1992). Les Interactions Verbales Ii. Paris: Armand Colin.

Kerbrat- Orecchioni, C. (1994). Les Interactions Verbales Iii. Paris: Armand Colin.

Leech, G. 1975. A Communicative Grammar Of English. London: Longman.

Leech, G. (1983). Principles Of Pragmatics. London New York: Longman.

Levinson, S. (1983). Pragmatics. Cambridge: Cambridge University Press.

Schiffrin, D. (1987). Discourse Markers. New York: Cambridge University Press.

Searle, J. R. (1969). Speech Acts: An Essay In The Philosophy Of Language. Cambridge: Cambridge

University Press.

Searle, J. R. (1979). Expression And Meaning- Studies In The Theory Of Speech Acts. Cambridge: Cambridge University Press.

Searle, J. R Et Al. (1980). Speech Act Theory And Pragmatics. Holland: D. Reidel.

Traverso, V. (2006). Des Echanges Ordinaires A Damas, Aspects De L'interaction En Arabe: Approche Comparative Et Interculturelle. Lyon Damas: Pul, Presses Universitaires De Lyon Ifpo, Institut Français Du Proche Orient.

Traverso, V. (2000). « Autour De La Mise En Œuvre D'une Comparaison Interculturelle », In V. Traverso (Ed.), Perspectives Interculturelles Sur L'interaction, Lyon : Pul.

Watts, R. J. (2003). Politeness- Key Topics In Sociolinguistics. New York: Cambridge University Press.

Yule, G. (1985). The Study Of Language. New York: Cambridge University Press.

Building comparable topical blog corpora for multiple languages

Andrew Salway
Uni Research, Bergen
andrew.salway@uni.no

Knut Hofland
Uni Research, Bergen
knut.hofland@uni.no

1 Introduction

This paper describes the construction of three corpora comprising English-language, French-language and Norwegian blogs related to the topic of climate change. Our approach may be applicable for creating other topically-focussed comparable web corpora in multiple languages.

As a site for large-scale and complex discourses about socially-relevant issues, the blogosphere is of increasing interest to social science research, e.g. to investigate how opinions are formed, how discourses are structured and evolve, and how different interest groups interact. The challenge then is to harvest all blogs related to a particular issue such as climate change which spans science, politics, social action, etc. and is discussed at international, national and local levels; its international dimension prompts interest in comparable multilingual corpora. Another important aspect of the blogosphere is the interaction between bloggers as evidenced by their linking patterns; thus blog corpora should contain data about hyperlinks as well as text.

Others have created blog corpora (e.g. Kehoe and Gee 2012). However, we are not aware of any corpus of "all" blogs related to a topic, nor any with hyperlink data, except our previous work (Salway at al. 2013). Previously our approach was to crawl from a small number of seed blogs. Whilst that was quite successful, it became apparent that it missed some important blogs and included some spurious ones. The main difference in the approach described here is that the burden of finding relevant blogs is placed more on web search engines that, we presume, are much better at indexing and crawling the web than we can hope to be.

2 Approach

Our working definition of a blog is a website created through one of a small set of blog authoring platforms: WordPress, Blogspot, Typepad and OverBlog. This restriction makes it feasible to write custom scripts to extract the main text of the post, comments, date, and different kinds of links. After extensive analysis we are confident that these platforms host the vast majority of blogs for our

topic and languages.

A set of potentially relevant blog posts was gathered using APIs to query three web search engines with a small set of core key terms for the topic, so as not to create imbalance between sub-topics and between different viewpoints. For English these terms were "climate change", "global warming" and "greenhouse effect"; these were translated into French (three terms with five inflections) and Norwegian (four terms with 12 inflections). Querying was done daily for 12 weeks and the rate of new posts (due to bloggers writing new posts and search engines re-ranking older ones) was monitored. Search engines limit the number of results returned, so after two weeks the set of query terms was expanded with frequent n-grams containing key terms, e.g. "of climate change".

About 170,000 blog posts were gathered. After manual inspection of data about blog posts containing certain numbers of key terms, it was decided to harvest all posts from blogs for which we had gathered 2 or more posts containing 2 or more instances of key terms. Data was also generated about hyperlinks in the 170,000 blog posts in order to check for blogs that were linked to from multiple blogs but that were missed in the previous step.

3 Current status

We have harvested all posts from 5563 English-language blogs, 2088 French-language blogs and 128 Norwegian blogs; approximately 9.7×10^6 blog posts and 5.9×10^9 words. Processing is underway to extract: the text content of each post, comments, date, data about article links (links from the main text of the post to any other site), blog roll links, and other links.

Next we will assess corpus quality, and clean where necessary, using techniques such as character and n-gram distribution to check topic, language and duplicates (cf. Biemann et al. 2013). Features particular to a blog corpus will also be checked, including the distribution of dates, and the network structure generated automatically from link data.

We expect that all processing and validation will be complete by July 2015. Our plan is to make the corpora available to researchers for download and for online analysis in the Corpuscle system[145].

Acknowledgements

This work is supported by the RCN's VERDIKT program. We are grateful to Dag Elgesem, Kjersti Fløttum, Anje Müller Gjesdal and Lubos Steskal for input on corpus design, and especially to Øystein Reigem for his work on normalisation, deduplication and link data extraction.

References

Biemann, C., Bildhauer, F., Evert, S., Goldhahn, D., Quasthoff, U., Schäfer, R., Simon, J., Swiezinski, L. and Zesch, T. 2013. "Scalable Construction of High-Quality Web Corpora". *Journal for Language Technology and Computational Linguistics* 28(2):23-59.

Kehoe, A. Gee, M. 2012. "Reader comments as an aboutness indicator in online texts: introducing the Birmingham Blog Corpus". *Studies in Variation, Contacts and Change in English* 12. Online at www.helsinki.fi/varieng/series/volumes/12/kehoe_gee/

Salway, A., Touileb, S. and Hofland, K. 2013. "Applying Corpus Techniques to Climate Change Blogs". In A. Hardie and R. Love (eds.) *Corpus Linguistics 2013 Abstract Book*. Available online at http://ucrel.lancs.ac.uk/cl2013/doc/CL2013-ABSTRACT-BOOK.pdf

[145] http://clarino.uib.no/korpuskel/page

Descriptive ethics on social media from the perspective of ideology as defined within systemic functional linguistics

Ramona Statche
University of Nottingham
Ramona.statche@nottingham.ac.uk

Svenja Adolphs
University of Nottingham
svenja.adolphs@nottingham.ac.uk

Chris James Carter
University of Nottingham
psxcc@nottingham.ac.uk

Ansgar Koene
University of Nottingham
ansgar.koene@nottingham.ac.uk

Derek McAuley
University of Nottingham
Derek.mcauley@nottingham.ac.uk

Claire O'Malley
University of Nottingham
Claire.omalley@nottingham.ac.uk

Elvira Perez
University of Nottingham
Elvira.perez@nottingham.ac.uk

Tom Rodden
University of Nottingham
Tom.rodden@nottingham.ac.uk

Issues of ethical behaviour are becoming increasingly important in discussions about social media, ranging from concerns about online behaviour and safety of individuals to debates about social media data usage in research, business and governance.

Within this context, the lack of a standardised vocabulary is a significant roadblock in the attempt to achieve a consensus on ethics. Vocabulary difficulties can be identified not only between separate domains of activity (compare, for example, legal language on data protection to language in various professional codes of practice) but often within the literature of a single domain. As an initial step towards addressing this gap, we propose a study of the fundamental meanings encoded through the terminology used in social media ethics. Can linguistic choices be a contributing factor in explaining observed conflicts between expressed attitudes and evidenced behaviour?

A descriptive ethics approach is applied to identify the ethical principles and choices evidenced in the codes of practice of a variety of UK chartered professional bodies. The same process is applied to the relevant European and UK legislation. Data is further enhanced by analysing the observed ethical choices made based on these regulations: legal verdicts, types of activities which received ethical approval, public reactions to evidenced practices etc.

The data identified through the descriptive ethics approach is then analysed in terms of meaning using the systemic functional linguistics framework. The context of the source texts is assessed to identify the ideological level employed.

It is shown how the ideological level of meaning of a vocabulary entry is perceived differently, based on the register variable of field which corresponds to different activity domains (legal, medical, technical etc.). Such fundamentally encoded differences at the variables of ideology and field lead to cases where the same vocabulary entry has several presumed definitions operating simultaneously. The resulting competing readings pose a great risk for miscommunication and further impede the already difficult task of reaching consensus on ethical perspectives.

Preliminary results indicate that vocabulary choices can be linked with behavioural outcomes. For example, 'privacy' is more tightly safeguarded through regulation when it is ideologically perceived as equating to a graded level of access and is a more easily dismissed concern when it is equated to secrecy.

Acknowledgement

This work forms part of the CaSMa project at the University of Nottingham, HORIZON Digital Economy Research institute, supported by ESRC grant ES/M00161X/1.

References

Martin, J. R. and White, P. R. R., 2005. *The Language of Evaluation. Appraisal in English.* Palgrave Macmillan.

Eggins, S., 2004. *An Introduction to Systemic Functional Linguistics*. 2nd ed. New York – London: Continuum.

Contrastive analysis "the Relative clauses based on Parallel corpus of Japanese and English "

Kazuko Tanabe
Japan Women's University
tanabeka@fc.jwu.ac.jp

This study is amied to contrast Japanese non- gap type relative clauses with their coresponding English translation based on Japanese –English News Article Alignment Data (Uchiyama and Isahara,2003) with *WebParaNews*, a searh engine developed by Chujo & Anthony, 2013).

This news paper corpus was composed by automatically aligning the articles of *Yomiuri news paper* and *The daily Yomiuri* from September 1989 to December 2001.

According to Comrie (1996, 1998, 1999, 2002), Japanese has the two kinds of attributive clause construction,which are almost equivalent to European case-gap relative clauses and also fact-S constructions (i.e. sentential complements with a nominal head) which Comrie called as "Asian-typed noun modification".

In my data searched on the Balanced Corpus of Contemporary Written Japanese (BCCWJ), in the case of the fact-S consturction, most of the nominal heads are the subjected two character Sino-Japanese. It is because in Japanese the subjected 2 character Sino-Japanese gerunds are frequently adopted when the abstract concepts are expressed.

Originated Japanese vocaburaly does not contain enoughwords to express the wide variety of the abstract concept. When this Fact-S construction are often used in order to explain the universal character or content of the abstract concept, Sino-Japanese gerunds are frequently employed.

As the result, it is clarified that in English translation, the abstract noun are hardly adoted and the significance expressed with verbal expressions.

Example:

(1) Jitai (situation)

Japanese:

Sengo seiji no ikizumari o shocho –suru jitai
Post war P ※dead lock ACC symbolize situation

da to ieyo.
copura Quo can be said

English:

'What happened in the Diet on Friday night symbolized a deadlock.'

If 'shocho-suru jitai' is directely translated it will be 'situation symbolizing (postwar)', however, 'jitai' is not translated here .

(2) hitsuyo (need)

Japanese:

Izon no sisei kara no dappi o
Reliance P pose from P casting off ACC

isogu hitsuyoo ga aru.
hurry need NOM exist

English:

'They must do away with their mentality of depending on the government.'

'Isogu hitsuyo' in the Japanese sentence seems to be translated into 'must' in English.

(3) hoshin(principle)

Japanese:

Senta de gennchi kunren o kaishi shi-tai hoshin
Center at training ACC start want principle

da.
copura

English:

'The ministry also intends to start training at the Japanese –made center.'

'Kaishi shi-tai hoshin' is presumed to be translated into 'intend to'.

In conclusion, Japanese fact-S construction does not tend to be translated into relative clause constraction in English. The English meaning of verbs or auxiliaries usually reflect the connotation of Japanese fact-S construction.

Selected learner errors in online writing and language aptitude

Sylwia Twardo
University of Warsaw
smtwardo@gmail.com

The aim of this paper is to analyse the possible connections between the results of a selected aspect of language aptitude (established by means of the FLAT-PL test, a Polish language version of the MLAT test) and the types and quantity of errors made by students when writing online at blended courses of English. In this study the analysis is focused on the results of the task Phonetic Script (Alfabet fonetyczny) and spelling errors. The FLAT-PL test was conducted on 71 students of the University of Warsaw who took part in blended courses of English at the CEFR levels B1, B2 and C1. The written texts were produced during the course of one semester and were different for each level. The output of respective students differed in quantity. The texts were extracted from the Moodle .mbz files with the use of Excel. The errors were coded and analysed by means of the AntConc concordancer. The correlations were calculated for the whole population and for each level separately and some statistically significant results were obtained.

References

Rysiewicz J., Foreign Language Aptitude Test—Polish (FLAT-PL), General characteristics, description, analysis, statistics and test administration procedures, Poznań 2011, retrieved on March 3rd, 2014, https://www.academia.edu/1744649/Foreign_Language_Aptitude_Test_-_Polish_FLAT-PL_Test_Uzdolnien_do_Nauki_Jezykow_Obcych_-_TUNJO_

The phraseology of the N that pattern in three discipline-specific pedagogic corpora

Benet Vincent
Coventry University
ab6667@coventry.ac.uk

1 Introduction

Pattern Grammar (Hunston & Francis, 2000) has helped advance research into phraseology by indicating that there are associations between complementation patterns and the meanings of the words that goven them. However, as Hunston (2011: 123) argues, patterns 'are often best seen as coming about because of a more pervasive phraseology than is represented by the pattern itself'. This observation raises the question of how one identifies such phraseologies.

The pattern that this study is interested in is is nouns followed by *that*-clauses (the **N that** pattern). This pattern has important to stance-construction in academic prose (Biber et al, 1999; Charles, 2007; Hunston, 2008). This importance is linked to the functions it performs in discourse, which, as Schmid (2000) shows, includes characterisation of propositions (as *facts, claims, ideas* etc.) and 'temporary concept formation'. Research into this pattern (e.g. Schmid 2000) has tended to focus on categorising nouns that occur in the pattern. Schmid's (2000) categorisation, shown in Table 1, takes account of the different relationship between the noun and the *that*-clause, which for the shaded meaning groups is appositive while for the other groups, it is modal in that the noun indicates how the proposition is to be interpreted.

Noun meaning group	Examples
Factual	*case, fact*
Linguistic	*assertion, claim, statement, suggestion*
Mental	*assumption, belief, discovery, idea*
Possible	*possibility, doubt*
Evidential	*evidence, proof, sign*
Emotional	*surprise, joy, fear*

Table 1: Meaning groups of N that nouns based on Schmid (2000)

This research has not generally concentrated on wider phraseological patterns associated with the **N that** pattern . An important exception to this is Hunston (2008), who investigated a number of **N that** nouns 'that best exemplify evaluation of

epistemic status' (Hunston 2008: 281) occurring in a corpus of texts from the New Scientist. Hunston found that these nouns participate in phraseologies that can be sorted into five main discourse functions:

- the idea, suggestion etc. exists
- the idea, suggestion etc. is evaluated
- the idea/suggestion etc. causes something
- the idea/suggestion etc. is caused by something
- the idea/suggestion is confirmed/disconfirmed

Previous studies have not, however, attempted a comprehensive investigation of phraseological patterning surrounding the **N that** pattern. Moreover, most of the research carried out so far has focused on large-scale corpora including a range of different registers; of the works cited above, only Charles (2007) has compared how the **N that** pattern varies in terms of its usage and frequency across subject-specific corpora. Pedagogic corpora (Willis, 2003), that is corpora composed entirely of texts that learners are exposed to in their learning context, are also much neglected in the literature. For these reasons, despite the clear pedagogical importance of research into the **N that** pattern (Charles, 2007), teachers and students of English for Academic Purposes (EAP) lack concrete information on which to make informed pedagogical decisions in an important area of understanding and creating authorial stance.

This study is an attempt to address the issues mentioned above by investigating broader patterns of use surrounding the **N that** pattern in corpora composed of core texts from three different disciplines studied by first year undergraduates at an English-medium university. In doing so, severl aims are pursued. The first of these is to investigate the extent to which instances of the **N that** pattern form part of frequent, identifiable 'semantic sequences' (Hunston, 2008). The second is to ascertain to what extent such patterning varies across the three subject-specific corpora investigated.

2 Corpora

The corpora used in this study (see Table 2) are small untagged subject-specific corpora composed of core texts from three compulsory first year undergraduate courses at a university in Turkey. The disciplines involved are Mathematics (Maths), Natural Sciences (NS) and Social and Political Science (SPS). These corpora therefore constitute what Willis (2003) terms 'pedagogic corpora' in that they consist of all (or nearly all) of the texts that the learners are exposed to during their studies.

Corpus	Word count
Mathematics	265,959
Natural Science (NS)	279,899
Social & Political Science (SPS)	280,095

Table 2: Corpora used in the study with word counts

As can be seen from Table 2, the corpora used in this study are relatively small. However, since they contain either the entire textbook – in the case of Mathematics – or a significant proportion of the extracts that students are expected to read while studying the SPS and NS courses, we can be fairly certain that findings based on these corpora are representative of the type of language that first year undergraduate students at the university will meet on these compulsory courses.

3 Method

Each corpus in turn was loaded into the *AntConc* freeware corpus software and a concordance search for *that* was undertaken. To make **N that** instances more salient, the item to the left of 'that' was sorted alphabetically so that, for example, all uninterrupted instances of *fact that* would be listed together. The resulting concordance lines were then saved as a text file so those which did not contain instances of **N that** could be removed. This search method allows those instances where the noun and the *that*-clause are separated by intervening material to be included (Charles, 2007).

These **N that** instances were then classified along the lines set out by Hunston (2008), that is, by grouping them according to similar discourse functions. In doing so, concepts such as semantic preference (Sinclair 2004) and 'long distance collocation (Siepman 2005) were particularly helpful. Moreover, Schmid's (2000) distinction between 'modal' and 'appositive' relationships was also used to distinguish instances.

4 Findings

As can be seen from Table 2 the number of instances of the **N that** pattern in the NS corpus normalised to hits per million is considerably higher than those found in the other corpora.

Corpus	Raw **N that** hits	**N that** hits (pmw)
Maths	181	681
NS	185	660
SPS	274	979

Table 2: **N that** hits in each corpus

The classification of **N that** instances yields further functions beyond the five proposed by Hunston

(2008). It is proposed to add to the list the functions:

- The fact etc… means/suggests that
- Definitions (e.g. *Kepler's 2nd Law is simply the observation that angular momentum is conserved in planetary motion*)

Modal **N that** instances are also treated separately.

Grouping the **N that** instances in this manner allows differences in terms of phraseology across the three corpora to be compared and to demonstrate how these disciplines evaluate knowledge claims. Each of these functions shows different distributions across the three disciplines with a degree of overlap. For example, modal **N that** instances relating to probability (*what is/find the probability that…*) are commonly found in Maths and NS corpora but less often in SPS; those relating to evidence are found in NS and, to a lesser extent in SPS. But the realisations of these apparently similar meanings differ across the corpora.

On the basis of regularities of meanings, it is possible to tentatively propose semantic sequences, some of which are relatively frequent, such as 'USE + *the fact that* [formula] + to find/calculate' in Maths. However, a number of these, while resonant with those indicated by Hunston (2008) are either quite infrequent or not clearly linked with a particular function. The variability of findings and fact that a significant percentage of instances in each of the corpora appear to be minority uses also raises the question of how such sequences might best be grouped.

The results therefore suggest that there are certain conventional sequences of meaning elements and that these are associated with particular contexts, but also that further research is needed with larger corpora to establish how they might best be distinguished.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. 1999. *Longman Grammar of Spoken and Written English.* Harlow: Longman.

Charles, M. 2007. "Argument or evidence: Disciplinary variation in the use of the Noun that pattern in stance construction." *English for Specific Purposes*, 26: 203–218

Hunston, S. 2008. "Starting with the small words: Patterns, lexis and semantic sequences." *International Journal of Corpus Linguistics* 13: 271-295

Hunston, S. 2011. *Corpus approaches to evaluation: phraseology and evaluative language*. New York: Routledge.

Hunston, S. and Francis, G. 2000. *Pattern grammar*. Amsterdam/Philadelphia: John Benjamins.

Schmid, H-J. 2000. *English abstract nouns as conceptual shells: from corpus to cognition*. Berlin/New York: Mouton de Gruyter.

Siepman, D. 2005. "Collocation, colligation and encoding dictionaries. Part I: Lexicographical aspects". *International Journal of Lexicography* 18 (4): 409-43

Sinclair, J. 2004. *Trust the text*. London: Routledge.

Willis, D. 2003. *Rules, patterns and words: grammar and lexis in English language teaching.* Cambridge: Cambridge University Press.

The representation of surveillance discourses in UK broadsheets: A corpus linguistic approach

Viola Wiegand
University of Nottingham
viola.wiegand@nottingham.ac.uk

In 2013 a worldwide public discussion on privacy and surveillance was triggered by the leak of confidential data from the US National Security Agency, alleging that the agency has been collecting massive amounts of emails and call metadata (Black 2013). Given the rapid worldwide expansion of surveillance measures since the September 11 ('9/11') terror attacks in 2001 (Lyon 2004), surveillance is arguably becoming a 'cultural keyword' (Williams 1983) and appears to gain great social significance. However, only relatively few studies have examined public discourses of surveillance from a linguistic perspective, with the notable exceptions of Barnard-Wills (2011) and MacDonald and Hunter (2013a; 2013b). Indeed, a corpus linguistic approach allows us to systematically analyse "the representation of social issues, global events or groups in society" (Mahlberg 2014: 220) by means of examining linguistic patterns. While some corpus linguists adopt concepts from Critical Discourse Analysis (e.g. Baker 2006; Baker et al. 2008), corpus linguistic investigations can also draw on theoretical models from other disciplines. For instance, McEnery (2009) has shown that 'key keywords' can be employed to test the validity of a sociological theory in a corpus. In a trial study I followed this approach by investigating the surveillance discourses in a news corpus according to a model of surveillance 'frames' (Barnard-Wills 2011). One frame that appeared particularly salient in the analysis of that study was a strong distinction between the terrorist 'them' and the righteous 'us' in order to legitimate surveillance. The results of this study thus suggested that the sampling period (newspaper articles from 2001 - 2005) and their connection to the 9/11 terror attacks had an impact on the surveillance discourses in the corpus. In the present study, this issue will be addressed more specifically by means of a diachronic comparison of surveillance discourses before and after 9/11 in order to explore the temporal dimension of surveillance. More generally, the present study is concerned with the various meanings of the term *surveillance* in news discourse. A related aim of the study is to further address the methodological research question of how the discursive representation of surveillance can be identified in a large newspaper corpus. Apart from contributing to the area of corpus linguistics, these findings are also expected to potentially benefit the growing interdisciplinary field of 'surveillance studies' (e.g. Zurawski 2007). The study will be based on a corpus of UK broadsheets covering the period from 1997 to 2005. This arrangement places the events of 9/11 in between two blocks of approximately equal duration, thus facilitating the comparison of 'pre- and post-9/11' surveillance discourses.

References

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, P., Gabrielatos, C., KhosraviNik, M., Kryzanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society, 19*(3), 273-306.

Barnard-Wills, D. (2011). UK news media discourses of surveillance. *The Sociological Quarterly, 52*(4), 548–567.

Black, I. (2013, June 10). NSA spying scandal: What we have learned. *The Guardian*. Retrieved from http://www.theguardian.com/world/2013/jun/10/nsa-spying-scandal-what-we-have-learned

Lyon, D. (2004). Globalizing surveillance: Comparative and sociological perspectives. *International Sociology, 19*(2), 135-149.

MacDonald, M. N., & Hunter, D. (2013a). The discourse of Olympic security: London 2012. *Discourse & Society, 24*(1), 66-88.

MacDonald, M. N., & Hunter, D. (2013b). Security, population and governmentality: UK counter-terrorism discourse (2007-2011). *Critical Approaches to Discourse Analysis Across Disciplines, 7*(1), 123-140.

Mahlberg, M. (2014). Corpus linguistics and discourse analysis. In K. P. Schneider & A. Barron (Eds.), *Pragmatics of Discourse* (pp. 215-238). Berlin: De Gruyter Mouton.

McEnery, T. (2009). Keywords and moral panics: Mary Whitehouse and media censorship. In D. Archer (Ed.), *What's in a Word-list? Investigating Word Frequency and Keyword Extraction* (pp. 93-124). Farnham: Ashgate.

Williams, R. (1983). *Keywords: A Vocabulary of Culture and Society* (2nd ed.). London: Fontana.

Zurawski, N. (2007). Einleitung [Introduction]. In N. Zurawski (Ed.), *Surveillance Studies: Perspektiven eines Forschungsfeldes* [Perspectives of a research field] (pp. 7-24). Opladen: Barbara Budrich.

Synthetism and analytism in the Celtic languages: Applying some newer typological indicators based on rank-frequency statistics

Andrew Wilson
Lancaster University
a.wilson@lancaster.ac.uk

Róisín Knight
Lancaster University
r.knight1@lancaster.ac.uk

1 Introduction

This study applies some newer quantitative typological indicators to elucidate relationships and evolution within the Celtic language family. These indicators are distinctive from earlier typological indicators (such as Greenberg's [1960] synthetism index) in that they require no morphosyntactic analysis but rely purely on rank-frequency or type-token statistics (Popescu & Altmann, 2008a, 2008b; Popescu, Mačutek & Altmann, 2009; Kelih, 2010). An important point about Greenberg's indices is that they require a fairly deep knowledge of the grammar in order to be applied reliably. Even with such knowledge, they involve substantial effort in manual analysis. Simpler indicators that can measure the same constructs are therefore to be welcomed. Descriptively, the work extends the typological analysis of Tristram (2009) on Celtic, which excluded three of the languages (Manx, Cornish, and Scottish Gaelic).

2 Theory

The power-law function for ranked word frequencies typically does not fit exactly and usually crosses the observed frequencies somewhere within the hapax legomena. Popescu and Altmann (2008a) have observed that, if the curve crosses the observed frequencies early, so that most of the hapax legomena lie above it, this indicates a tendency towards synthetism; however, if the curve crosses the observed frequencies late, so that most of the hapax legomena lie below it, then this indicates a tendency towards analytism. This is because analytic languages tend to use the same word-form multiple times whilst synthetic languages use a greater number of unique forms (because the lexeme changes form to signal grammatical information). In the former case, the function underestimates the hapax legomena and, in the latter case, it overestimates them.

Kelih (2010) has also suggested that type-token statistics alone might be an indicator of typology, without needing to fit the power function. This is because an increase in the number of hapax legomena – the main underlying feature of the rank-frequency-based indicators – necessarily leads to a change in the type-token relationship overall.

3 Data

The data for this pilot study is a small translation corpus of ten Psalms per language, giving 70 texts in total. All of the Celtic languages are included: Welsh, Cornish, and Breton (the "P-Celtic" branch); and Manx, Scottish Gaelic, and Irish (the "Q-Celtic" branch). Two periods of Irish are included as separate samples. Each text was processed individually.

4 Results

For the languages where comparative data are available (Welsh, Breton, and Irish), all of the rank-frequency-based indicators are rank-order identical with Greenberg's synthetism index, as computed by Tristram (2009). Such a direct comparison has not previously been made for any language, and this finding bodes well for future applications of these indicators.

More concretely, the indicators demonstrate not only that Irish has evolved from a greater to a lesser degree of synthetism but also that synthetic versus analytic tendencies within Celtic seem not to be linked in any way to the ancestral Q- versus P-Celtic classification. This picture was not entirely clear in Tristram's (2009) study, since she did not compute Greenberg's index for Manx, Cornish, and Scottish Gaelic. In our study, Manx (a Q-Celtic language) is the most analytic of all; in contrast, Cornish (a P-Celtic language) is the second most synthetic language, more so than Modern Irish (Q-Celtic). Since the diachronic tendency in most European languages has been a move away from synthetism, it seems unlikely that disparities in text dates lie behind these results: the Cornish texts are the most recent, whilst the Manx texts only post-date the Early Modern Irish texts by around a century.

The type-token statistics tell a slightly different story, so it has to be assumed that the two approaches are actually not directly comparable. In this case, the pattern is more directly suggestive of Q- versus P-Celtic relations, with the two historical stages of Irish particularly close to one another.

5 Conclusion

This research, despite drawing only on a small pilot sample of Psalm texts, and with limitations on text dates, suggests that the newer typological indicators may be of considerable value in investigating morphosyntactic typological variation. As far as Celtic is concerned, our continuing work drawing on other discourse types and other dates will surely tell

an interesting story.

References

Greenberg, J.H. 1960. "A quantitative approach to the morphological typology of languages". *International Journal of American Linguistics,* 26: 178-194.

Kelih, E. 2010. "The type-token relationship in Slavic parallel texts". *Glottometrics,* 20: 1-11.

Popescu, I.-I. and Altmann, G. 2008a. "Hapax legomena and language typology". *Journal of Quantitative Linguistics,* 15(4): 370-378.

Popescu, I.-I. and Altmann, G. 2008b. "Zipf's mean and language typology". *Glottometrics,* 16: 31-37.

Popescu, I.-I., Mačutek, J. and Altmann, G. 2009. *Aspects of word frequencies.* Lüdenscheid: RAM-Verlag.

Tristram, H.L.C. 2009. "Wie weit sind die inselkeltischen Sprachen (und das Englische) analytisiert?" In U. Hinrichs, N. Reiter and S. Tornow (Eds.), *Eurolinguistik: Entwicklung und Perspektiven* (pp. 255-280). Wiesbaden: Harrassowitz.

Conflicting news discourse of political protests: a corpus-based cognitive approach to CDA

May L-Y Wong
The University of Hong Kong
maylywong@hku.hk

1 Introduction and research methodology

This study uses methods from corpus linguistics and theoretical constructs from cognitive linguistics to examine patterns of representation around Occupy Central, a recent political protest in Hong Kong, in two corpora of English-language newspaper articles published in *China Daily* and the *South China Morning Post* (SCMP). Using the online corpus analysis tool *Wmatrix* (Rayson 2008), an analysis of key semantic domains of the press reports enabled a comparison of the two newspapers according to the following thematic categories: group/organisation, confrontation, characterising attributes and consequences, all of which relate to different aspects of the protest. The analysis subsequently considered three discursive strategies, namely structural configuration, framing and identification (Hart 2013 a/b, 2014a/b), that are mediated through conceptualisations that representations in text evoke.

2 Results and discussion

Given the opposing political stances of the two newspapers under consideration, we should expect to find subtle differences in linguistic representations which reflect these conflicting ideological positions. In accordance with findings from quantitative corpus analysis, it has been found that the news reports of *China Daily* invoke conceptualisations of police as a homogeneous organised entity of professionals where police involvement is legitimated to restore public order in response to the actions of the protesters. When it comes to how issues of public order are reported in news discourse, much of previous research has proved that the conservative press tend to favour representations which refrain from challenging dominant power relations and instead preserve the social status quo (e.g. Fowler 1991; Montgomery 1986; Trew 1979). However, in the *South China Morning Post* which is more liberal in its orientation, representations in texts serve to invoke conceptualisations of police, rather than protestors, as instigators of forceful actions. The conceptual patterns upheld in each paper are thus reflective of what van Dijk (1998) refers to as an 'ideological square' – a structure of mutual opposition involving simultaneous positive Self-representation and

negative Other-representation. Consistent with the contrasting political stances of the two newspapers, then, the government and its authorities such as the police force and the court are aligned with the Self in *China Daily* and legitimated while protesters are positioned as Other and delegitimated. The converse is seen in SCMP where the protests are legitimated and the police response delegitimated. These differences, therefore, represent the events in ideologically different ways and serve to apportion blame and agency along alternative lines commensurate with institutional stances and identities.

3 Conclusion

Hopefully, my analysis has shown that integrating critical discourse analysis (CDA) with corpus linguistics and cognitive linguistics in a more balanced way has the potential to identify typical linguistic patterns across many thousands of words as well as reveal underlying construal operations which fulfill an ideological potential in media discourse. The aim of this approach is then to demonstrate that systematic, unbiased and scientifically grounded critical discourse research is perfectly possible when equipped with the right tools and theories of language. Corpus linguistic techniques can be viewed as an additional methodological tool that can be combined with CDA approaches to text analysis in order to "reach a set of more wide-reaching, representative and objective conclusions" (Baker and McEnery 2014: 479). However, by the use of a cognitive linguistic approach to CDA, quantitative investigations of corpus examples could be focussed and contextualised in such a way that particular linguistic instantiations in discourse can be further analysed in relation to a wider range of conceptual phenomena which may carry some ideological load. It would then be reasonable to view that interfacing between qualitative research afforded by cognitive linguistics and quantitative corpus methods is necessary to provide a 'full' critical discourse analysis where both qualitative and quantitative analyses become mutually reinforcing and enriching.

References

Baker, P. and McEnery, T. 2014. "'Find the doctors of death': press representation of foreign doctors working in the NHS, a corpus-based approach". In A. Jaworski and N. Coupland (eds.) *The discourse reader* (3rd ed.) (pp. 465-480). London and New York: Routledge.

Fowler, R. 1991. *Language in the news: discourse and ideology in the press*. London: Routledge.

Hart, C. 2013a. "Constructing contexts through grammar: cognitive models and conceptualisation in British newspaper reports of political protests". In J. Flowerdew (ed.) *Discourse and contexts: contemporary applied linguistics*, vol. 3 (pp. 159-184). London: Continuum.

Hart, C. 2013b. "Event-construal in press reports of violence in two recent political protests: a cognitive linguistic approach to CDA". *Journal of Language and Politics* 12(3):400-423.

Hart, C. 2014a. "Construal operations in online press reports of political protests". In C. Hart and P. Cap (eds.) *Contemporary critical discourse studies* (pp. 167-188). London: Bloomsbury.

Hart, C. 2014b. *Discourse, grammar and ideology: functional and cognitive perspectives*. London: Bloomsbury.

Montgomery, M. 1986. *An introduction to language and society*. London: Routledge.

Rayson, P. 2008. "From key words to key semantic domains". *International Journal of Corpus Linguistics* 13(4):519-549.

Trew, T. 1979. "Theory and ideology at work". In R. Fowler, B. Hodge and G. Kress (eds.) *Language and control* (pp. 94-116). London, Boston and Henley: Routledge & Keegan Paul.

Van Dijk, T. 1998. *Ideology: a multidisciplinary approach*. London: Sage.

Automatic Analysis and Modelling for Dialogue Translation Based on Parallel Corpus

Xiaojun Zhang
Dublin City University
Xzhang@computing.dcu.ie

Longyue Wang
Dublin City University
Vincentwang0229@gmail.com

Qun Liu
Dublin City University
Qliu@computing.dcu.ie

Discourse analysis oriented to dialogue-text machine translation (MT) system differs in many respects from ordinary text MT system, as it is subject to a number of specific constraints, both of a technical and an interactional nature. Here, we mention three. First of all, the texts of dialogue contain a lot of ill-formed phenomenon including repetition, ellipsis, disorder, and broken sentences (utterances). Secondly, dialogue involving computers is usually highly restricted in domain. In addition to these, parallel dialogue texts are poor-resourced and insufficiently-collected to train an MT system.

To analysize the dialogue-text for spoken language translation (SLP) system, a Chinese-English parallel dialogue corpus is essentially required for the task of dialog machine translation. The corpus contains around 2 million of sentence pairs, which are acquired from parallel dialogue texts such as movie subtitles, multi-person conversations and social media. The crowd-sourced translation of movies/episodes is a vast resource to harvest parallel dialogue texts besides European Parliament Interpreting Corpus (EPIC) [146], and Spoken BNC2014[147]. However, the issue of crowd-sourced translation quality arises. We need automatically estimate the quality of each translated film dialogue when we decide to collect it. The corpus will be annotated with meta data including language pair (English into Chinese, Chinese into English), genre (science fiction, action-adventure, animation, etc.), parallel titles, writer(s)/author(s), translator(s) as well as responsible party (the person who sample the data etc.). In addition, we extract and label the 'domains' of some dialogues in film, i.e., we label some dialogue with 'AIRPORT' when they happened at airport scenes of the films, and with 'EMERGENCY' when they happen in the situation of dangers or risks. The parallel corpus is aligned at sentence level and all plain texts are formatted in TMX with extensible markup language (XML) and encoded in UTF-8.

Next step of dialogue analysis is to segment the texts in corpus into discourse segments based on the above parallel dialogue corpus in order to meet the requirements of dialogue machine translation. Each segment should be a coherent discourse unit which is independent with each other. Research into discourse analysis technologies based on parallel discourse corpus, including anaphora, co-reference, ellipsis, speech act etc.

One linguistic feature of anaphor resolution we mentioned is that the dropped pronouns occur frequently in Chinese whilst seldom in English. English is a non-pro-drop language (Haspelmath, 2001) whilst Chinese is a pro-drop language (Huang, 1989) which subject in a sentence is always optional, especially, in the dialogue. That is, pronouns as the anaphors are always dropped in the source language and we should complete them in the target language. We aim to develop an adapted methodology to cross the barrier ofpronominalized collocation via pronoun resolution in dialogue translation from pro-drop source languages (such as Chinese, Persian etc.) into non-pro-drop target languages (such as English).We may retrieve many null matches of subjects from large-scale phrase-aligned bilingual parallel corpus used in machine translation, which we can trace to the dropped pronoun antecedents. The definite and indefinite articles in English also provide us a clue to analyze the anaphoric dependency in Chinese. We constructed a pronoun-tagged Chinese/English bilingual dialogue corpus to identify the dropped pronouns in Chinese with the help of their English translations. An example of identified dropped pronouns in italic type in the dialogue from the movie *Crouching Tiger, Hidden Dragon* as:

玉娇龙: 这剑套真好看.
俞秀莲: 再好看也是凶器. 刃上染了血, 就不会说它好看了.
玉娇龙: 在江湖上走来走去的是不是很好玩?
俞秀莲: 走江湖, 靠得是人熟, 讲信, 讲义. 不讲信义, 可就玩不长了.

Master Long: The scabbard is so beautiful.
Governor Yu: *It*'s beautiful but dangerous. Once *you* see *it* tainted with blood, its beauty is hard to admire.
Master Long: *It* must be exciting to ba a fighter, to be totally free!
Governor Yu: *Fighters* have rules too: friendship, trust, integrity. Without rules, *we* wouldn't survive for long.

Furthermore, the factors such as speakers and intentions also affect the SLT system. By investigating the influence of the above phenomena on dialog translation, new machine translation

[146] http://sslmitdev-online.sslmit.unibo.it/corpora/corpora.php
[147] http://cass.lancs.ac.uk/?page_id=1386

models will be established to incorporate these factors to improve the accuracy of dialog translation.

Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of the ADAPT Centre (www.adaptcentrel.ie) at Dublin City University, Ireland. It is also supported by the HUAWEI TECHNOLOGIES Co., LTD and National Social Science Foundation of China (Grant 10CYY006) as part of Shaanxi Normal University, China.

References

Haspelmath, M. 2001. "The European linguistic area: standard average European". *Language Typology and Language Universals* (1):1492-1510.

Huang, C.T. 1989. "Pro-drop in Chinese: A Generalized Control Theory". In Jaeggli, O. and K. J. Safir (eds.) *The Null Subject Parameter*. London: Kluwer Academic Publisher.

Absence of Prepositions in Time Adverbials: Comparison of '*day' tokens in Brown and LOB corpora

Shunji Yamazaki
Daito Bunka University

yamazaki@ic.daito.ac.jp

As noted by Quirk *et al* (1985: 692), prepositions introducing some time adverbials may be absent, so that "the time adverbial takes the form of a noun phrase instead of a prepositional phrase". Such variation can alternatively be described as variable overt vs. zero-marking of the adverbial by a preposition, e.g. *I'll see you (on) Monday*. There are linguistic conditions under which this variation is limited: e.g. Quirk *et al* (1985: 692), Quirk and Greenbaum, (1973: 156), and Celce-Murcia and Larsen-Freeman (1999: 403) point out that prepositions of time adverbials are always absent immediately before the deictic words *last, next, this, that* (e.g. *I met Mr. Leech last Sunday*), and before the quantitative words *some* and *every* (e.g. *Every summer they go back to their home town*).

Under other conditions, such as when the adverbial denotes delimited periods of time including years, months, weeks, or days of the week, both alternatives are possible. However, several researchers have suggested that there may be context- or dialect-sensitive variation in their frequencies of use. Sonoda (2002: 19) comments that there is some conditioning by formality level: "*on* and *for* are omitted most frequently in informal styles". Algeo (1988: 14) states that with such (named) periods of time, "the omitted preposition is Common English", but that there are several areas of difference between British and American English: in some cases, British English "has no preposition, but one would be expected in American" English, and by contrast "British [English] usually requires a preposition (*on*) with days of the week, whereas American [English] can have the preposition or omit it".

The present research compares data from the Brown and LOB corpora to examine the following dimensions of variation in omission of adverbial prepositions:

- Variation by preposition (some prepositions may be more likely to be omitted than others: for example, as a function of overall preposition frequency, or (conversely) specificity of meaning).

- Variation by regional differences (prepositions are more omitted in American English than British English).
- Variation with sentence position of the adverbial (preposition omission may be expected with higher frequency in sentence-initial adverbials than in sentence-final adverbials).
- Variation by genre (in particular, more formal genres should more often favour overt markers).
- Variation by semantic relationship between the sentence and the adverbial (preposition omission should be more frequent with more general time adverbial meanings, e.g. with expressions of time duration).
- Variation by lexical item and/or frequency (some frequent expressions may favour preposition omission).

References

Algeo, John (1988). British and American grammatical differences. *International Journal of Lexicography* 1: 1-31.

Celce-Murcia, M. and D. Larsen-Freeman. (1999). *The Grammar Book*. Boston: Heinle-Heinle.

Quirk, R. and S. Greenbaum. (1973). *A Univeristy Grammar of English*. Essex: Longman.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Sonoda, Kenji. (2002). Omission of Prepositions in Time Adverbials in Present-day Spoken AmE. *Bulletin of Nagasaki University School of Health Sciences*. 15(2):19-25.

Corpus of Russian Student Texts: goals, error classification and annotation

Natalia Zevakhina
National Research University Higher School of Economics
natalia.zevakhina@gmail.com

Svetlana Dzhakupova
National Research University Higher School of Economics
Svetlanads@yandex.ru

Elmira Mustakimova
National Research University Higher School of Economics
egmustakimova_2@edu.hse.ru

1 Introduction

The Corpus of Russian Student Texts (CoRST)148 is collected at the Linguistic Laboratory for Corpus Technologies at the National Research University Higher School of Economics, Moscow. The project started in 2013, and since then the size of corpus has reached about 2 650 000 tokens. The part of corpus containing more than 300 000 of tokens is error annotated: in total, about 8 500 errors are tagged.

The corpus comprises students' written texts. Despite that students are Russian native speakers, the texts contain a considerable amount of fragments which we regard as linguistic deviations, or errors. Following the developers of learner corpora, we mark them up. Hereby, we pursue both pedagogical and research aims.

The pedagogical aims are similar to those of learner corpus: we encourage our students to avoid the ways of expressing their ideas that do not correspond to the norms of academic writing. The research aims are based on the assumption that errors are markers of language change (Rakhilina 2014; Glovinskaya 2010 among others). Indeed, if most speakers systematically make the same so-called "error", we observe a consistent language trend, which may possibly shape a new linguistic norm in the future.

Types of annotation

The texts of CoRST are supplemented with metalinguistic, morphological and error markup.

Metalinguistic markup contains information about a text (type of a text, year, semester/module) and its author (age, gender, first language, region of residency, faculty/department, year of studying, bachelor/master, academic major). The corpus

[148] http://web-corpora.net/CoRST/

includes the following types of texts: course paper, abstract, essay, etc. The corpus includes texts written by the students of the different academic majors.

Morphological markup is carried out automatically with help of the morphological analyzer MyStem (Segalovich and Titov 1997-2014). However, morphological ambiguity is not resolved: every ambiguous word is provided with all possible grammatical analyses. The tag set of 52 morphological labels meets the standard established by RNC.

In error markup, we follow the principle of multilayered annotation. First, apart from tags that classify error types, we also introduce tags that classify the cause of an error. Second, a text fragment that contains an error may be corrected in different ways and, consequently, it may have several different error tags; in such cases, all possible tags are provided. The corpus comprises 20 higher level tags and 19 lower level tags (in total, 39 tags). For example, "lex" is a label for lexical errors; this is a higher level tag which includes 4 lower level tags: wrong word, wrong phrase, metonymy, and intensifier (see Error Classification below). We also marked the beginning and end of each citation in order to rule out the fragments, which are not authored by a student, from the morphological search. The error annotation is carried out manually using the interface of Les Crocodiles 2.6 (Arkhangelsky 2012).

2 Error classification

Error markup in CoRST is based on the following error classification: lexical, grammatical, discourse, and stylistic errors. Apart from the latter type, all other types of errors correspond to the traditional language levels: lexicon, grammar and discourse.

Main types of lexical errors are lexical errors in a narrow sense (word, phrase, intensifier or metonymy error), word formation errors (including paronyms and aspectual errors), errors in nominalizations and auxiliary verbs.

Among grammatical errors, we identify the following: agreement, government and coordination errors, errors in comparative and superlative constructions, errors in complex sentences (including errors in sentential arguments and relative clauses), errors in the choice of conjunctions, coreference violations (including errors in pronouns and in converbs), errors in nominal and verbal inflection, omissions (including ellipsis), construction violations.

To the discourse errors, we attribute the following cases: meta-textual comments, mixing of direct and indirect speech, incoherent sentences, parcellation (division of sentences into incomplete units), logical errors, wrong use of linking words, wrong word order, tautology, inappropriate topicalization.

Under stylistic errors, we mean any mismatch between the style of a text and its type, including inappropriate use of colloquial or official style.

As for the causes of mistakes, we annotate two types of causes: typo and construction blending. The latter seems to be important since discussion of errors makes sense only with respect to particular constructions.

The tag set does not contain labels for orthographic or punctuation errors. They surely may be found in student texts; however, such mistakes seem to be less relevant for linguistic research.

Finally, it is worth noting that while developing the annotation system, we tried to balance theoretical views on errors classification and practical purposes, i.e., annotation convenience.

Acknowledgments

The results of the project "Corpus studies of language variation: from deviations to linguistic norm", carried out within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) in 2015, are presented in this work.

References

Arkhangelsky T. 2012. *Les Crocodiles 2.6*. [Software]. Moscow.

Glovinskaya, M. 2000. Aktivnye processy v grammatike // *Russkiy yazyk kontsa XX stoletiya*. M. (Active processes in grammar // *The Russian language in the end of the 20th century*. Moscow).

Rakhilina, E. 2014. Stepeni sravneniya v svete russkoy grammatiki oshibok // *Sbornik k 10-letiyu NRC*. M. (Comparative degrees in the view of error Russian grammar // *Collection towards the tenth anniversary of NRC*. Moscow).

Segalovich I., Titov V. 1997-2014. *MyStem*. [Software]. Available from https://tech.yandex.ru/mystem/

Corpus-based approach for analysis of the structure of static visual narratives

Dace Znotiņa
University of Latvia
Daceznotina@gmail.com

Inga Znotiņa
Liepaja University, Ventspils University College
inga.s.znotina@gmail.com

1 The structure of static visual narratives

Narrative is "a mentally produced organization largely (..) dependent upon the cognizing activities of an experiential or perceiving subject" (Flanagan 2008) that can be externally represented in different media (e.g., texts, films, pictures). Therefore, static visual stimuli (pictures) can be narrative if the mental model that is made during comprehension of stimuli contains elements corresponding with at least following components: (a) two or more temporally or causally linked events represented and/or inferred (Herman 2009), (b) specific characters, (c) fictional world, (d) time, (e) immersion into the story world (Sanford and Emmott 2012). In order to access and analyze the mental model of narrative, it must be presented in an external form of narrative – pictures or texts.

2 Use of corpus-based approach

The use of corpus-based approach for analysis of visual material and structural analysis is unusual because visual material is not textual and linear and can be ambiguously interpreted. In order to use corpus linguistics tools, a special data set of texts must be made. The process of transforming visual stimuli into a proper textual format includes 4 steps: (1) subject verbally describes presented picture (using Concurrent Think-Aloud Protocol) supplying textual representation of the picture; (2) spoken text is transcribed into machine-readable format; (3) text is coded with content analysis using categories that are derived from elements of narrative structure; (4) text can be formalized or annotated using these categories. After these four steps, initial visual material is now linear, sequential, machine-readable, and contains all necessary additional structural information.

To analyze narrative structure, two ways of transforming spoken texts can be used — annotation and formalization.

One of the approaches for annotation is use of problem-oriented tagging (McEnery and Wilson 2001). The set of tags can be derived from the categories of content analysis that are based on the elements of the narrative structure. It can be done by using XML. The annotated texts can be analyzed using any software that can be used for XML annotated corpora. This analysis allows to obtain frequencies of specific narrative elements or to collect content of any specific structural entities. Further structural analysis can be done, if the software can process collocations between annotated elements. For example, it can answer what elements tend to stick together and other structural inquiries. The main advantage of using annotations is a possibility to preserve the content of the text within the structure. That, in turn, allows using concordances to access the content of spoken text. The main disadvantage is the need for a specialized software.

The other way of creating usable material is via formalization (schematic representation). In this case, the corresponding textual fragment can be substituted with a formal denotation (word, symbol, or abbreviation) that represents the element of narrative structure. These symbols can be derived from the same categories of content analysis. The output texts from formalization process are in a form of sentence and can be analyzed with any corpus research software because any formal symbol can be regarded as a word. These formal sentences can be structurally analyzed using word frequencies and collocations. The possibility to use any corpus linguistics software is the main advantage for this approach. The main disadvantage is the lack of immediate connection between structure and its initial content.

3 Current research

In the current research, 50 pictures (static, single-frame) were used – photography and drawings (documental, art). 20 participants described them verbally producing 1000 short texts. Content analysis (categories: character, event/action, time, space, world knowledge, emotion/immersion) of these texts took 4 previously described steps (~6 hours for each 20 minutes of spoken text). The results from corpus analysis contain the frequencies of narrative elements and the relations between them (the mental structure of the static visual stimuli).

4 Conclusions

In the case when there are only a few texts to be analyzed, there is no need for using corpus methods because of the manual processes required for transforming the material into the proper format. But in a case when there is a larger amount of texts that are similar in their length, content and structure, the quantitative approach can be efficient to observe some universal patterns and tendencies. Structural analysis might require the first three steps of

transformation process regardless whether a corpus is built, and the last step of transformation (annotation or formalization) can be done automatically. If that is the case, use of the corpus-based approach can offer a faster and more precise way for obtaining empirically valid results.

Acknowledgements

This work was partly funded by European Social Fund, project "Doktora studiju attīstība Liepājas Universitātē" (grant No.2009 / 0127 / 1DP / 1.1.2.1.2. / 09 / IPIA / VIAA / 018).

References

Flanagan, J. 2008. *Knowing More Than We Can Tell: The Cognitive Structure of Narrative Comprehension.* In Partial Answers: Journal of Literature and the History of Ideas, Volume 6, Number 2, June 2008: 323-245.

Herman, D. 2009. *Basic Elements of Narrative*. Oxford: Wiley-Blackwell.

McEnery, T., Wilson, A. 2001. *Corpus Linguistics. An Introduction.* Second Edition. Edinburgh: Edinburgh University Press.

Sanford, A.J. and Emmott, C. 2012. *Mind, Brain and Narrative.* Cambridge: Cambridge University Press.

Learner corpus *Esam*: a new corpus for researching Baltic interlanguage

Inga Znotiņa
Liepaja University,Ventspils University College
inga.s.znotina@gmail.com

1 Introduction

The aim of the present paper is to describe a new publicly accessible learner corpus *Esam*[149] which is being built as a part of the author's ongoing PhD research. The corpus is made to investigate Baltic interlanguage – the interlanguage that forms when a person with the background of one Baltic language (Latvian or Lithuanian) learns the second Baltic language.

2 Design of the corpus

The corpus consists of texts that have been written by university students, learners of the second Baltic language; namely, Latvian for students of Lithuanian background, and Lithuanian for students of Latvian background. The texts are written independently on a variety of topics: "My family and friends", "The place where I would like to return", "A strange day in my life", etc. Students who wanted to write on their own topics were encouraged to do so. Each text was written as a homework, so students had access to their notebooks, dictionaries and study materials. All currently collected texts were written between 2007 and 2014. The length of the texts varies from 45 to 500 words.

3 Data collection

The texts included in the corpus were collected without any specific goal by the teachers of the second Baltic language in the respective universities over several years. They were then given to the creator of the corpus who tracked down the authors and asked them for permission to use the texts.

Those authors who agreed to allow using their texts in the corpus, signed a permission which, among other things, states that:

- the texts included in the corpus can be made publicly accessible;
- the identities of the authors are not to be revealed anywhere except the list of authors on the website of the corpus (if the author agrees to be included in it).

[149] The name of the corpus was chosen to emphasize the closeness of both Baltic languages – Latvian and Lithuanian. *Esam* means 'we are' in Latvian, as well as in colloquial Lithuanian.

Therefore, all texts undergo the process of anonymization before getting included in the corpus. It means that all the information that can reveal the author's identity is either removed (such places are tagged with the tag <izlaid> in the text) or replaced to retain the integrity of the text (such places are tagged with the tag pair <anon> </anon>).

4 Markup and annotation

So far, the texts included in the corpus have not been marked in any way apart from anonymization. However, the corpus's website offers a table which includes metainformation about each text which could later be turned into markup. This information includes:

- the code of the text (allowing to identify specific text files of the corpus);
- the code of the author (allowing to identify several texts written by the same author);
- the topic of the text;
- the amount of words in the text (counted before anonymization);
- the language of the text;
- the semester in which the respective text was written (namely, first or second consecutive semester of learning the language);
- language of instruction.

The texts have also not been annotated yet.

5 Access to the corpus

A sample of the corpus is currently publicly accessible on *esamcorpus.wordpress.com*. It is downloadable as a collection of *.txt files which can then be researched with any software that supports this file type. The files have been tested to work with Anthony Lawrence's *AntConc*[150].

The size of the sample is about 15,000 words, and it consists of 68 texts. All texts included in the sample have been anonymized. The sample only consists of texts in Lithuanian that were written by Latvian students.

6 Future plans

The size of the corpus already collected with permissions exceeds 40,000 words, and it will all become publicly available once it is anonymized. The collaborating teachers are still collecting texts from current students, so the size of the corpus might increase in the future.

Markup and annotation of the corpus is also planned. Markup is expected to include the aforementioned metainformation categories. Annotation should include error annotation, part-of speech annotation, and syntactic sentence type annotation.

All changes and additions to the corpus are described in the *News* section of the corpus's website.

Acknowledgements

The creator of the corpus would like to thank the authors for allowing to use their texts in research. Special thanks to all the second Baltic language teachers who are helping to gather materials and create the corpus.

[150] Freely available on http://www.laurenceanthony.net/software/antconc/