Crowdsourcing Formulaic Phrases: towards a new type of spoken corpus

Adolphs, S., Knight, D. Smith, C. and Price, D.

Corpora have revolutionised the way we describe and analyse language in use. The sheer scale of collections of texts, along with the appropriate software for structuring and analysing this data, has led to a fuller understanding of the characteristics of language use in context. However, the development of corpora has been unbalanced. The assembly of collections of written texts is relatively straightforward, and as a result, the field has a number of very large corpora which focus on mainly written texts, although often with some spoken elements included, e.g. the *COCA* (520m words), *GloWbE* (1.9b), and the *enTenTen* (19b). In addition, a number of corpora now include samples of language used in social media and other web contexts alongside more traditional written and (transcribed) spoken language samples, e.g. the *Open American National Corpus* (planned corpus size 100m words, mirroring the *British National Corpus*). Conversely, the development of spoken corpora has lagged behind, mainly due to the time-consuming nature of recording and transcribing spoken content. Most of the spoken corpora that exist consist of material that is easily gathered by automated collection software, such as radio talk show and television news transcripts and other entertainment programming (e.g. the spoken elements of *COCA*). The nature of this spoken discourse is described as unscripted, however, it is certainly constrained, e.g. talk show radio has certain expectations about how the host will moderate the discussion. While the scripted/constrained oral content in these spoken corpora has proved informative in terns of the nature of spoken discourse (see Adolphs and Carter, 2013; Raso and Mello, 2014; Aijmer, 2002 and Carter and McCarthy, 1999 for notable studies), it is no substitute for spontaneous, unscripted oral discourse. Furthermore, even the automated collection of scripted/constrained spoken discourse has not yet enabled the development of large spoken corpora of a size comparable to the largest written corpora (e.g. the spoken component of the 100m *British National Corpus* is only 10m words with a further 10m words added in the new spoken BNC2014). The 10m word subcorpus of the BNC contains 4m words of spontaneous speech, and is controlled for a number of sociolinguistic and contextual variables. There are a number of smaller spoken corpora available, e.g. the *Michigan Corpus of Academic Spoken English* (MICASE) which at just under 2m words is both modest in size and quite specialised in content. This trend is reflected in other corpora of spoken discourse.

Spontaneous spoken discourse forms a large part of everyday language use, and the development of larger and more representative corpora of spontaneous oral language is therefore desirable to inform linguistic description. The main constraint to this ambition has always been the time-consuming nature and financial cost related to the compilation of such corpora. Spoken corpora provide a unique resource for the exploration of how people interact in real-life communicative contexts. Depending on how spoken corpora are annotated (as discussed below), they present opportunities for examining patterns in, for example, spoken lexis and grammar, pragmatics, dialect and language variation. Spoken corpora are now used in a variety of different fields from translation to reference and grammar works, to studies of language change.

The need for spontaneous unscripted corpora seems uncontroversial, however, compiling such corpora in the traditional way remains a formidable task. Advances have been made in other areas utilizing the power of people volunteering information about what they think and do. This approach is often referred to as *crowdsourcing*, and it holds the promise to

both overcome some of the difficulties outlined above, and to add useful aspects to corpus compilation which traditional methods cannot offer.

This paper thus explores a new approach to collecting samples of naturally occurring spoken language samples, which may allow researchers to take advantage of the burgeoning area of information crowdsourcing. Instead of relying on the typical recording and transcribing of spoken discourse, crowdsourcing may allow the collection of real-time data 'in the wild' by having participants report the language they hear around them. Specifically, we aim to investigate the level of precision and recall of the 'crowd' when it comes to reporting language they have heard in real certain contexts, alongside the use of a crowdsourcing toolkit to facilitate this task. This method of 'reporting' usage does come with its own issues of course, many of which have been highlighted in the literature on Discourse Completion Tasks (Schauer and Adolphs, 2006), and can merely be regarded as a proxy for usage. Investigating user memory in this context can therefore only be regarded as a first step in assessing the overall viability of the proposed approach to collecting language samples. As a focusing device for selection of reported language samples, we draw on the use of formulaic phrases, an area that have received considerable attention from different areas in applied linguistics.

What is crowdsourcing?

Crowdsourcing is an approach that involves the outs*ourcing* of specific forms of tasks or activities via open calls to a large network of unknown labourers (i.e. the *crowd*). While the term *crowdsourcing* was only coined in 2005, early incarnations of this approach can be traced back long before this date. The earliest example was during the development of the 1857 edition of the New Oxford English Dictionary, with the editor calling for the British public to submit words and examples to be included in the dictionary. Driven by the tagline 'anyone can help',[1] this call resulted in over 6 million submissions to the dictionary, a number that would have been impossible for the editor and his team to produce alone. While this early approach was clearly time-intensive, given that it took over 70 years to amass this dataset, the initiative was certainly impressive in terms of the scale and ingenuity.

The onset of the digital age has resulted in an exponential growth in the development and utility of more advanced crowdsourcing methods, which have the potential to target a more extensive, global online 'crowd'. One of the earliest examples is Amazon's mechanical Turk (*MTurk*, 2005);[2] an online, digital crowdsourcing utility. *MTurk* was originally created to aid the development and validation of language resources. It focuses specifically on distributing Human Intelligence Tasks (HITs) to the crowd. Anonymous volunteers (referred to by serial numbers alone) are paid small amounts of money upon the completion of tasks, providing a cost- and time-effective way of getting work done. The human input is particularly useful for tasks that computers are unable to carry out solely on artificial intelligence, and so require real-world involvement by human informants.

Despite the usefulness of crowdsourcing in certain contexts, there are a number of potential pitfalls and challenges to note, especially when it comes to using crowdsourcing to build a collection of language samples. A crowdsourcing approach relies heavily on the availability of technology to the crowd. In the study described here, we issued users with devices and an interface ready to record language samples. However, in order to scale up the experiment, it would be necessary to rely on people's own mobile devices and compatibility with the software. The other challenge of using crowdsourcing over a corpus-building approach

[1] http://blog.oxforddictionaries.com/2014/02/can-world-englishes-benefit-crowdsourcing/
[2] www.mturk.com/mturk/welcome

is that careful sociolinguistic sampling is more difficult with this approach. Given that much of the research into spoken language based on spoken corpora relies on careful sampling this is a potential issue. The decision of using a crowdsourcing approach to collecting spoken data in the way we describe here is therefore dependent on considering a number of trade-offs involved, with the main one being the quality of reported versus 'used' data and careful sampling of contextual variables versus access to a potentially larger set of different contexts of use and ease of data collection. Our paper will explore the feasibility of a crowdsourcing approach in the first instance, however, these other factors would need to be taken into account in any larger scale study.

Crowdsourcing in the public domain

Crowdsourcing approaches are now widely used across a range of different public-facing online and app-based platforms as a means of completing numerous different tasks and activities. Crowdsourcing is being used to gather information about local history and historical events,[3] as well as for data mining and solving data problems.[4] Crowdsourcing sites are also used to raise money for charities where individuals pledge to, for example, participate in sporting and other events,[5] or raise awareness of particular healthcare charities through the growth of moustaches.[6] Further to this, 2015 witnessed a particular surge in the use of crowdsourcing methods in the context of investment and start-up companies. The principle aim of these sites is for the crowd to raise money to help fund projects, products, and businesses (an initiative now known as 'crowdfunding'), with some of the most popular websites including *Kickstarter*,[7] *CrowdCube*[8] and *StartUpValley*.[9]

With fees as small as 5p paid for a 20 minute task for some of the HIT-based online crowdsourcing systems (such as *MTurk*), questions have been raised as to whether this approach is exploitative and simply represents a digital version of slave labour (an online sweat shop). To explore the question of exploitation, Ipeirotis (2010) surveyed 1,000 random Turkers to get a clearer sense on the types of people who are using HIT systems and why they are doing so. In terms of understanding 'who' is engaging in these activities, 47% of the respondents questioned were found to be US citizens, 34% were from India, and the remaining 19% were from 66 other countries. 50% of all respondents had university level education. Only 15% of the US Turkers said that they used the site for primary income purposes, whilst this was true for 27% of the Indian Turkers. Far more common reasons for using *MTurk* was as a way of spending free time fruitfully (70% US, 60% India), and as a secondary source of income (60% US, 37% India). Respondents also noted that they completed tasks because they viewed them as being 'fun' (40% US, 27% India). The results of this study suggest that people use *MTurk* to make some extra money (or as a hobby), so while there are serious concerns about potentially bypassing workers' rights as well as about the quality of the data collected in this way, it may not be as exploitative as it first appears (see Adda et al., 2014: 1 for further discussion on this).

The notion of tasks being 'fun' lies at the heart of the development of many forms of crowdsourcing activities, e.g. *Games With A Purpose* (*GWAP* – see Adda et al., 2014: 3). *GWAP*'s are web and app-based systems that turn the process of crowdsourcing into a more competitive, game-based activity. For example, they can encourage the participant to score as many points as possible, or beat an opponent in a certain length of time in a bid to be crowned

[3] http://crowdsourced.micropasts.org/
[4] www.crowdflower.com
[5] E.g. www.justgiving.com and www.gofundme.com
[6] https://uk.movember.com/
[7] https://www.kickstarter.com/
[8] https://www.crowdcube.com/
[9] http://www.startupvalley.com/

the victor. Examples of *GWAP*'s include Ahn's (2006) *ESP game*, which crowdsourced the labelling of photos and images, a system later licensed to Google, forming the foundations of the *Google Image Labeler*. Chamberlain et al.'s *Phrase Detectives* (2008)[10] is a further example of a *GWAP* that focuses on crowdsourcing the annotation of language data. While participants can be incentivised to play *GWAP*'s with some offering digital credit or a free entry to a competition to win a cash or other prize, the premise behind many *GWAP*'s is not always to earn money, differentiating them from *MTurk* and other traditional HIT-based systems.

Crowdsourcing for academic research

Within the academic domain, crowdsourcing methods have been used extensively in the 'hard sciences' (in areas such as engineering, biology and Natural Language Processing), and are increasingly being applied to projects in arts, humanities and social sciences. Online platforms that facilitate 'people-powered research' include www.zooniverse.org which brings together a range of different research-based crowdsourcing sites that aim to recruit participants to carry out a range of tasks from the classification of galaxies[11] to the transcription of handwritten documents by Shakespeare[12] and the works of Jeremy Bentham.[13]

In the applied linguistic context, crowdsourcing methods have been used for the translation and/or transcription of speech to enhance speech recognition systems (see work by Gelas et al., 2011; Evanini et al., 2010; Marge et al., 2010; McGraw et al., 2010; Novotney and Callison-Burch, 2010 and Callison-Burch, 2009), and to annotate datasets (Asheghi et al., 2014).

Crowdsourcing has also been used in the collection of spoken linguistic data. One example is the compilation of the online *Speech Accent Archive.*[14] In another, Goldman et al. (2013) built the bespoke *Dialäkt* and *Voice Äpp* as a means for crowdsourcing data that will help identify and differentiate Swiss German dialects in locations around Switzerland. To contribute data using the *Dialäkt Äpp,* participants are given a list of 15 different words (taken from the *Sprachatlas der Deutschen Schweiz*), each of the which has 5 different localised pronunciations affixed to it, in both written and spoken (audio) form. Participants are required to define which of the pronunciations appears most like their own (i.e. dialectal variant choice). *Voice Äpp*, currently in development, is a more advanced version of *Dialäkt Äpp*, which asks users to pronounce individual words. It then uses speech recognition techniques to identify the variants and it localizes the user through a process of geotagging via their phones. *Voice Äpp* thus aims to provide more accurate/fine-grained voice profiling information about articulation, speech rate, and other oral characteristics through production-based tasks. To date, 39,168 participants (with a 42:58 female-male ratio) have contributed data to the project via this app, although only 4% of all downloads have led to a complete recording of all words.

Hughes et al.'s (2010) *Datahound* is another crowdsourcing app for the collection of spoken audio recordings, comprising of text-based prompts from common web queries that are presented on the screen of the app. The app provides users with textual prompts to read out which are recorded and time and date stamped automatically before being uploaded to a central server. Text-based metadata is also collected for each contributor, adding information about the gender, age and accent of the user, along with basic details of the 'acoustic environment' in which the data was recorded (e.g. whether it was indoors, outdoors or in a car). A similar community-driven approach is utilized in a data collection app being used in the development

[10] https://anawiki.essex.ac.uk/phrasedetectives
[11] www.galaxyzoo.org/
[12] http://www.shakespearesworld.org
[13] http://blogs.ucl.ac.uk/transcribe-bentham/
[14] http://accent.gmu.edu

of the *CorCenCC* (Corpws Cenedlaethol Cymraeg Cyfoes (The National Corpus of Contemporary Welsh)), a corpus construction project that began in 2016.[15]

The use of crowdsourcing has been useful for the collection of spoken data focusing on oral aspects ranging from accent to dialect to capturing the complete heritage of a language. This suggests it might also be feasible to use it in the compilation of corpora focusing on spontaneous unscripted spoken language.

Formulaic phrases in spoken language

One of the key corpus insights into language in use (including spoken language) is that much of language is formulaic in nature. For example, it is possible to say many different things to dismiss a sales assistant (e.g. *Your assistance is not required, I don't need any help, Go away!*), but a common and preferred formulation is *(I'm) just looking, thanks*. This confers the meaning of declining help in an expected way which is considered appropriate and polite. The more common a speech act or language function is in the real world (e.g. apologizing, requesting, offering sympathy), the more likely there are conventionalized phrases available to handle that situation. This makes knowledge of these recurrent phrases key to communicating effectively in discourse. There has been much work in describing the characteristics of formulaic sequences (e.g. Schmitt, 2004; Wray, 2002; Nattinger and DeCarrico, 1992 and Pawley and Syder, 1983), but the field is still at a nascent stage in its capability to identify and list the most important sequences for research and language pedagogy. Lists of various types of formulaic language have been developed, including the *PHRASE List* (Martinez and Schmitt, 2012), the *Academic Formulas List* (Simpson-Vlach and Ellis, 2010) and the *Academic Collocation List* (Ackermann and Chen, 2013). However, these lists inevitably reflect the corpora they were compiled from. Since none of the source corpora had major proportions of unscripted general spoken discourse, it is unclear how well the lists represent the formulaic sequences used in everyday spontaneous spoken English.

Crowdsourcing may be useful in compiling the most important phrases used for everyday situations. But this would only be possible if participants are able to recall phrases that they hear in particular contexts, close to the events themselves, so that they can report them accurately. There is little research in the crowdsourcing literature which addresses the ability to identify and recall phrases appropriate to particular situations, and the main goal of this study is to explore this issue.

Quality control of crowdsourced data

The ability to reliably and accurately recall and report formulaic sequences is a quality control issue of particular interest for the present study. But quality control is of wider interest in crowdsourcing methodology. It is important to note that in the case of *MTurk* and many of the other crowdsourcing utilities mentioned thus far, while some training is required, participants do not tend to require *specialist* training to complete their tasks. Given this, and the fact that participants are typically anonymous, quality-related safeguards need to be put into place prior to their distribution.

Adda et al. (2010) emphasises the need for crowdsourcing tasks to be broken down into their most basic form to be effective, as the more advanced tasks become, the more 'risk' they carry. In addition, Callison-Burch and Dredze (2010) emphasise the need to provide clear and coherent instructions for anything that is crowdsourced (i.e. instructions should be pitched at non-experts). Researchers should pilot the tasks they create to make sure they are achievable before the tasks are released to the general public. For quality control purposes, it is also important that systems are put in place which allow researchers to effectively determine 'good'

[15] http://www.corcencc.org/

contributions and to screen out 'bad' ones with tests for accuracy carried out when processing the results of the data.

If these quality-related considerations are adhered to, crowdsourcing methods can arguably produce accurate results that can prove valuable for future research. Evanini et al.'s (2010) study examined the accuracy of *MTurk* transcription and discovered that it was, in fact, close to typical human annotation, containing a low rate of additional errors (a finding supported in the work of Callison-Burch (2009)). Similarly, when testing the accuracy of corrected automatic speech recognisers (carried out using crowdsourcing methods) Gruenstien et al. (2009) and McGraw et al. (2009) both found that near-expert results were obtained, on the condition that the appropriate data and instructions were provided to the Turkers at the beginning of the task. But the accuracy and completion rates may well depend on the language being researched, as less-widely-spoken languages could be more prone to error simply because there is a smaller crowd to access for research purposes.

In sum, crowdsourcing has been used successfully in many types research, including in applied linguistics, and may be part of the solution to the limitations of spoken corpus compilation. But this would only be the case if participants can reliably and accurately report the formulaic sequences they use as they move through a number of different contexts throughout their day. This study explores the viability of crowdsourcing for corpus construction by exploring the ability of people to remember formulaic sequences they have heard, and to report them accurately. Our context is that of an academic lecture, leading to the following questions: Can participants recall the formulaic sequences they have heard in an academic lecture? Furthermore, we explore whether participants are able to report the formulaic sequences they heard without erroneously including other formulaic sequences which were not in the lecture.

Methodology

Ahead of the main study reported in this paper, a small-scale pilot study was first devised to determine whether the idea of crowdsourcing formulaic sequences is worth pursuing, and to underpin the design of the main study.

Target formulaic sequences

A university lecture in business studies (see below) was analysed for formulaic sequences, and twenty-one sequences were selected from it. These sequences occurred in the lecture between one and five times, and had the range of frequencies in general spoken academic discourse between 0-349 as indicated by frequency figures from the British Academic Spoken English corpus (*BASE*). Some sequences focused on business content (*decision making, systems control*), while others served more general discourse functions (*at the same time* = concurrent timing). We also made a list of nine formulaic sequences which did not occur in the lecture, but were similar in kind to the target items (hereafter referred to as *decoy phrases*). The complete list of targets and distractors can be found in Appendix 1.

Participants

For our pilot study, five participants were selected from the postgraduate student population of the School of English at the University of Nottingham including two native (NS) and three non-native (NNS) speakers of English. Three of the students were engaged in research towards doctorate degrees and two were studying at Masters level. All non-native speakers had achieved IELTS scores of 6.5 or above in exams taken up to four years prior to this study and their first languages were either Thai or Chinese.

Procedure

The aim of the pilot study was to understand whether participants were able to recall formulaic sequences that they had previously heard in a particular discourse context. The context we chose was an academic lecture. An 18-minute section of a business studies lecture was selected from the Nottingham Multi-Modal corpus. It was selected as being a reasonably self-contained information unit and a practical length for the experiments. The five participants watched the video together in a seminar room. They were asked to approach the task as they would any lecture at the university. They were allowed to take notes. Once the video had finished, smart phones were handed out and once shown how to log in, participants were asked to follow the instructions on the screen to complete the crowdsourcing task. In the crowdsourcing task, participants were asked to confirm whether or not they had heard the 30 formulaic sequences presented on a list displayed on their phone. (The list included the 21 targets which did occur on the lecture and 9 distractors.) Once the participants had finished the recall task, the smart phones were collected and the questionnaire was handed out for completion. The questionnaire had three sections. The first asked for basic participant data such as what course they were studying, whether they were a native speaker of English and their English language qualifications. The second section focussed on the task of identifying phrases included the level of confidence in selecting phrases heard and how they approached the task of selecting phrases. The final section asked about their thoughts on the crowdsourcing software used. The full text of the questionnaire is shown in Appendix 2.

After all participants had completed the questionnaire they were informed of the purpose of the study. As this was a pilot study the participants were also asked about the way the experiment was conducted, in particular whether they would have been happier being interviewed than completing the questionnaire. Only two of the participants responded to this question, one preferring an interview and another saying they would prefer the questionnaire. In the main study, the questionnaire was retained but was adapted slightly following our experience of the pilot study (see below).

The crowdsourcing toolkit

The crowd-sourcing tasks were used as exemplar tasks to drive the development of a reusable, multi-purpose crowd-sourcing toolkit. This toolkit focused on the concept of crowd-sourcing activities as *campaigns.* A campaign is defined as a set of crowd-sourcing tasks to be performed bounded by time. A *campaign administrator* is able to create a new campaign, define a series of tasks/activities within that campaign, and set a time period in which the campaign is to run. *Campaign participants* select to participate in one or more campaigns. Further, participants are able to register the types of campaigns that they are interested in to aid in the recruitment to crowd-sourcing activities by campaign administrators. The prototype toolkit was developed as a Microsoft Azure cloud service with a web-based user-interface (including a mobile optimised site) and an application programming interface for integration with other applications. For the purposes of this study, the developer was responsible for creating the interfaces to the crowd-sourcing activities (multi-choice questionnaire etc) for expediency and testing purposes. However, the toolkit was developed with the intention that administrators should be able to create their own interfaces and share them with other administrators, thus creating a *marketplace* for crowd-sourcing.

Results

The participants' answers were then analysed to determine the rate at which they were able to recall and report the target items from the lecture, and also to reject the decoy phrases.

Recall and Precision

Participants' ability to accurately report the formulaic sequences from the lecture can be conceptualized in two ways. The first way entails the number of sequences in the lecture which the participants were able to recall and report. We will refer to this as *recall*. It was calculated as a proportion of the recalled sequences divided by the total number of phrases (21) in the lecture:

- Recall = number of reported sequences which occurred in the lecture / 21

The second conceptualization involves the accuracy of the reporting. Crowdsourcing is of little value if participants erroneously report language which in fact they did not hear. We refer to this as *precision*. It was calculated by dividing reported sequences which actually occurred in the lecture by the total sequences reported (i.e. reported actual sequences + reported decoys). This measure takes account of the number of decoy phrases erroneously reported.

- Precision = number of reported sequences which occurred in the lecture / number of actual sequences reported + number of decoy phrases reported

The results of a combined recall and precision analysis can be seen in Figure *1*.

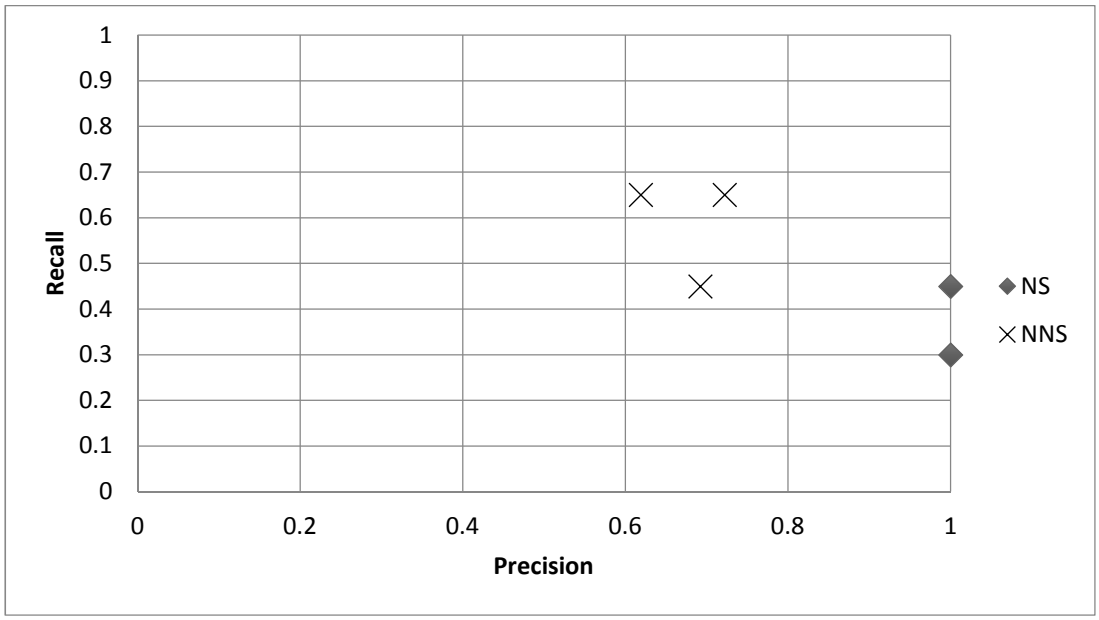

Figure 1: Scatter plot of recall against precision for sequences reported

The graph shows that the five participants were able to recall a reasonable number of sequences from the lecture, ranging from about 30% to 70%. Furthermore, the participants had good precision, with around 70% or more of the sequences selected occurring in the lecture. There

seems to be a trade-off between recall and precision, with higher recall percentages relating to lower precision percentages, and vice versa. It is also notable that the two participants with the highest precision (100%) were the two native speakers, with all of their reported sequences actually occurring in the lecture.

Viability of the crowdsourcing elicitation procedure

Results from the questionnaire showed the participants were happy with the recall task, and could understand both the instructions and the requirements of the study. Both native speakers reported that they found the recall task difficult, whereas all three of the non-native speakers reported that they found the task okay (Question 4). All but one said they were only partially confident that the phrases they selected were present in the lecture with the remaining participants saying they were fully confident (Question 5). Question 6 on the pilot study questionnaire asked the participants if they were more confident about some phrases they selected being actually present in the lecture than others; all participants answered yes to this question. It then asked them to list the phrases they were most confident about and then those they were least confident about. This part of the question did not capture the information we were hoping to gather. Many of the phrases listed had not been selected by the participant or were not even on the original list of phrases they were shown. Impressions of the crowdsourcing application itself were generally positive. When asked how easy the software was to use (Question 8), three participants said it was 'Easy', one said 'Very Easy', and one said 'Okay'.

Overall, the pilot results suggest that participants may well be able to supply useful information about discourse they have heard, thus confirming the viability of this approach. The pilot study also helped in refining the design of the main study. A larger-scale study was then carried out with the aim of testing a larger sample with a refined methodology. Given that there appears to be some difference between the reporting of native and non-native speakers, the subsequent study will need to include both of these participant groups.

Crowdsourcing Main Study

Methodology

The same basic methodology used in the pilot study was used in the final experiment, but with a few revisions based on the pilot study results. The same target and decoy words were used, the same lecture input, and the same smartphone elicitation procedure. The questionnaire was revised slightly. Question 6 that was not successful in the pilot (see above) and was removed.

We added one new element to the research design. The 21 formulaic sequences on our list were only a sample of those occurring in the lecture, and it is quite possible that participants might be able to recall and report sequences they had heard which were not on the list. We therefore asked the participants to add any phrases they recalled from the lecture that were not presented on our list. This was prompted in two places in the study. Firstly, there was the option to add an extra sequence on the smart phone application. This was then followed up with an additional question in the questionnaire asking participants to note down any sequences they recalled that were not on the list supplied on the handset. The complete text of the revised questionnaire is available in Appendix 3. This data should provide an enhanced indication of the participants' overall ability to recall the complete range of formulaic sequences in the lecture.

Participants

The study included 40 participants. They consisted of 24 native speakers and 16 non-native speakers from the undergraduate and postgraduate community of the University of Nottingham. Participants volunteered in response to an email sent to students across the university with the exception of those in the Business school. The majority of participants involved were postgraduates; only seven of the 40 were undergraduates. The non-native speakers had a wide range of first languages and English language qualifications. All had IELTS scores of 6.0 or higher or TOEFL IBT of 90 or higher.

Results

Recall and Precision

The recall and precision proportions are illustrated in Figure 2. Two extreme outliers (one native and one non-native speaker) have been excluded from this analysis.

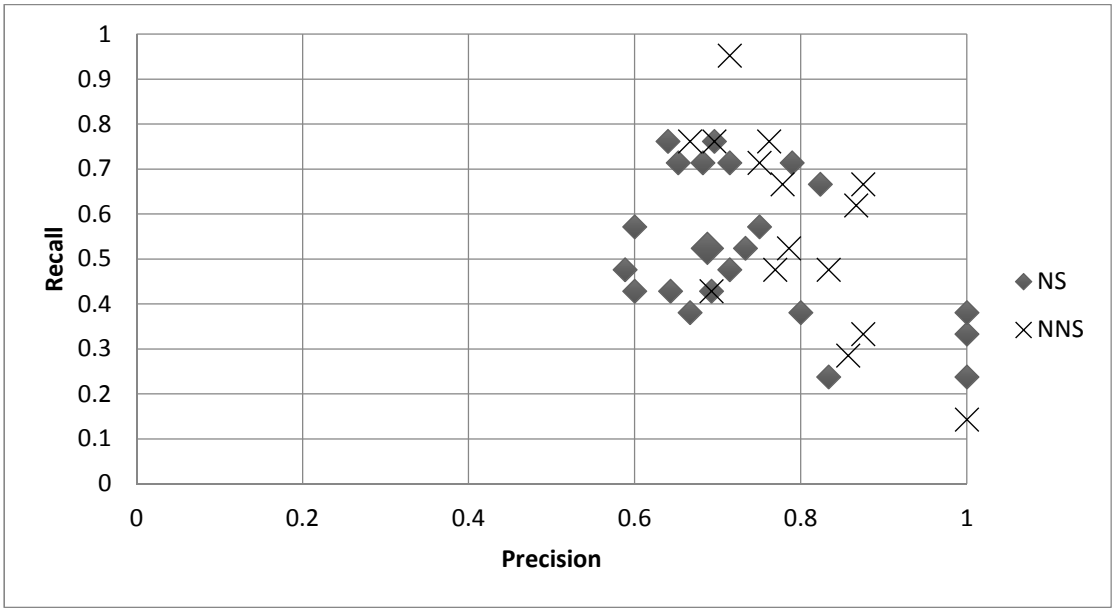

Figure 2: Scatter plot of recall against precision for sequences reported (the larger diamond shows two participants that shared the same precision and recall scores).

As with the pilot study, participants were able to recall and report a substantial number of the formulaic sequences occurring in the lecture, generally between 40%-80% of the 21 target sequences. Fourteen participants (35%; 14/40) were able to report 60% or more of the target sequences on the list, although only one was able to report more than 90%. However the graph also shows a great deal of variability in the recall proportions, with 20% (8 participants) reporting 40% or fewer target sequences.

The precision figures are also similar to the pilot results, with virtually all participants (with one near miss) achieving 60% or better accuracy. Just as with recall, there is evidence of considerable variation across participants. It is noteworthy that four participants achieved perfect precision (all reported sequences did actually occur in the lecture). Looking at the graph as a whole, the same inverse relationship between recall and precision that appeared in the pilot study also appears here. The participants who recalled more target items tended to report larger numbers of decoy phrases as well. Conversely, participants with the best precision tended to recall fewer target sequences. This is to be expected: the more precise participants only reported

sequences if they were very sure of them. Less precise participants may have been more willing to report sequences even if they had a much less certain sense of having heard them.

When the participants were asked about their confidence levels in selecting phrases in the questionnaire, most stated that they were only partially confident that the formulaic sequences they selected were actually heard in the lecture. The six participants who said they were fully confident, did have precision scores towards the higher end of the scale (all over 0.75), however, many of the other participants also had scores in this range, so it appears that the reported level of confidence cannot be used as a reliable predictor of precision. This suggests that future crowdsourcing may have to contain specific instructions. If the researcher wishes to gather the maximum number of language items heard in discourse, then asking participants to report what they are reasonably confident they have heard may be enough. Of course, we did not research how various instructions would affect participant behaviour, so this would be a useful point for future research.

Unlike in the pilot study, there does not seem to be any major differences between the precision and recall of native and non-native speakers in this larger experiment. To illustrate this more clearly, we have separated the native and non-native results in Figure 3 (native speakers) and Figure 4 (non-native speakers). The recall results of the native speakers cluster between 40% and 80%, and we find the same clustering for non-native speakers. The native English speakers seem to cluster somewhat more tightly, but this is mainly because there were more native participants than non-native ones.

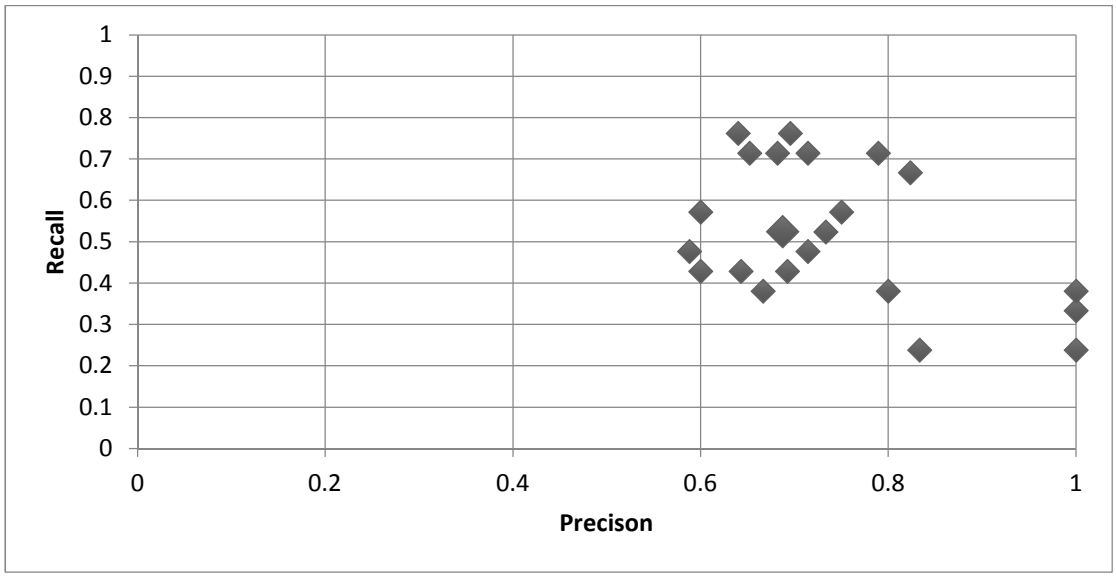

Figure 3: Scatter plot of recall against precision for sequences reported (natives)

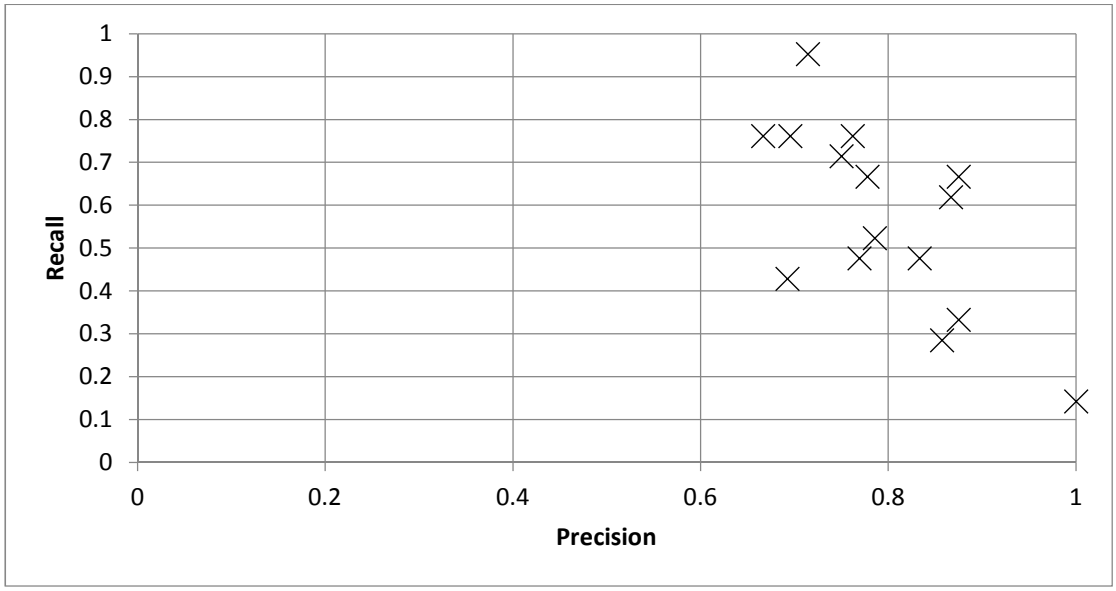

Figure 4: Scatter plot of recall against precision for sequences reported (non-natives)

Formulaic sequences reported in addition to those on the list

The use of a list allowed us to explore the reporting of formulaic sequences in a controlled manner, including exploring how many decoy phrases were reported. However, the real advantage of crowdsourcing would be to gather linguistic aspects with no prompting. To explore this, we asked participants to report sequences in addition to those that were on the list. In terms of elicitation methods, the questionnaire was much more effective than the smartphones. With the exception of one participant, all participants added more sequences on the questionnaire than on the handset. 19 participants added no sequences on smartphone, but when prompted again on the questionnaire, did write at least one sequence.

The reluctance to add sequences on the smartphone has strong implications for crowdsourcing 'in the wild' where people's smartphones would be the main form of data capture. In the questionnaire, 22 (55%) people reported that they were not familiar with a touch screen, or not being used to manipulating one and preferring a phone with a keypad. This proved a problem in this study, but smartphones are now completely dominating the market, with over a 75% share in 2015 compared to 'dumb' phones.[16] This means unfamiliarity with smart phones will almost certainly become less of an issue as times move on. Also, if people use their own devices (rather than those supplied by researchers), this may also help mitigate some of the reluctance found in our study. If apps such as ours are used in future research, improvements to the user interface will probably be needed to help increase users' confidence with the system, and perhaps encourage more sequences to be reported when the app is used over a longer period of time.

Out of the 40 participants, 37 (92.5%) offered at least one additional sequence. (The three participants who did not add any phrases either on the handset or on the questionnaire are omitted from the following analysis.) In total, 212 sequences were added. Of these, 80

[16] https://www.comscore.com/Insights/Rankings/comScore-Reports-May-2015-US-Smartphone-Subscriber-Market-Share

sequences were added on the handsets (although 53 of those were added by the same person)[17] and 132 were added on the questionnaire. This amounts to 181 unique sequences. Only 13 of the phrases were added by more than one person, the remaining 168 phrases were added only once. The most people to add the same phrase was five. The median number of sequences added per person was 3.5. The phrases added are of course only useful language if they were in fact present in the lecture which is addressed in the following analysis.

The recall measure, used for the analysis of selected phrases, is not appropriate here as we could never expect a participant to recall all of the words used in the lecture. Precision however is a very important measure as it allows us to see how well the phrases supplied by the participants relate to the actual language to which they were exposed. The precision for the added formulaic sequences is calculated by dividing the number of correctly reported sequences (those actually in the lecture) by the total number of phrases supplied by the participant (the correctly reported phrases plus the erroneous phrases that were not in the lecture). Figure 5 shows a comparison of the precision figures for target sequences selected from our list (*selected*) and added phrases (*added*). For this analysis, we looked at only exact matches with the ngrams extracted from the lecture. The fact that the majority of the participants fall in the top left quarter illustrates the difference in precision between selected and added phrases. With selected phrases this is nearly always above 0.60 (60%), but with added phrases it is typically under this figure. Perhaps the precision of selected items is not surprising, as the form of the items were given and it only needed to be recognized. But with the added sequences, the participants needed to reproduce the form themselves, and they were largely unable to do this, at least not always completely accurately.

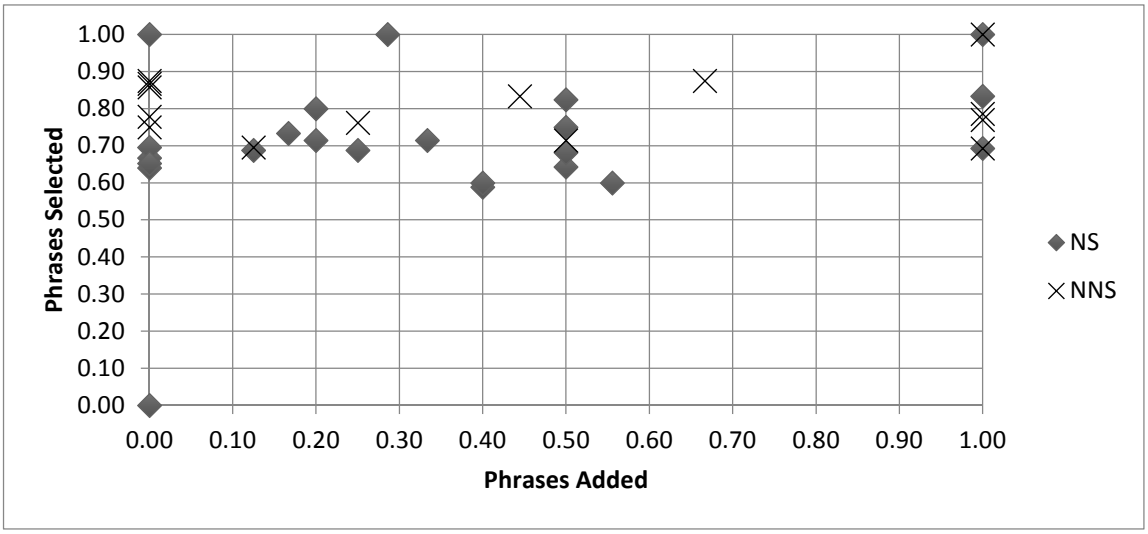

Figure 5: Comparison of precision in selected and added phrases (Exact match)

A closer examination of the phrases showed that many of the sequences added by participants were very similar, but not identical to phrases found in the lecture. Our second analysis allowed phrases that were very similar to those found in the lecture to be counted as matches. The Damerau-Levenshtein algorithm was used to find the closest match to the ngrams

[17] This particular individual (along with two others) completed the handset part of the task twice due to the application crashing the first time. The data from both submissions, however, was successfully logged to the database. It was decided to use the first submission from each person for the phrase selection task but to combine the phrases added as the phrases were not the same each time.

in the text. The resulting pairs of phrases were analysed and several categories of changes became apparent:

- Grammatical changes of number, tense etc.
- Use of abbreviations
- Substitution of different prepositions
- Substitution of similar sounding words
- Shared concepts expressed differently

Of these categories, 'grammatical changes' and the 'use of abbreviations' resulted in phrases that were the most similar to those actually found in the lecture. Examples of such pairs of phrases are shown in Table 1. These categories contained a total of 15 unique phrases and 18 instances of these phrases. Figure 6 shows the precision result where sequences such as these are considered as correct. As we would expect, there is a slight shift towards the right of the graph with the precision of more of the participants creeping over the 50% mark, but the majority of participants still fall below this. It thus appears that even grammatical-leniency does not help participants to report heard phrases with good precision.

Phrase Added	**Closest Match in Transcript**
strategy planning dept	strategic planning departments
military strategy military strategist	military strategists
system control view	systems control view
focus on individuals	focussing on individuals
follow it	followed it

Table 1: Examples of phrases with grammatical changes or the use of abbreviations

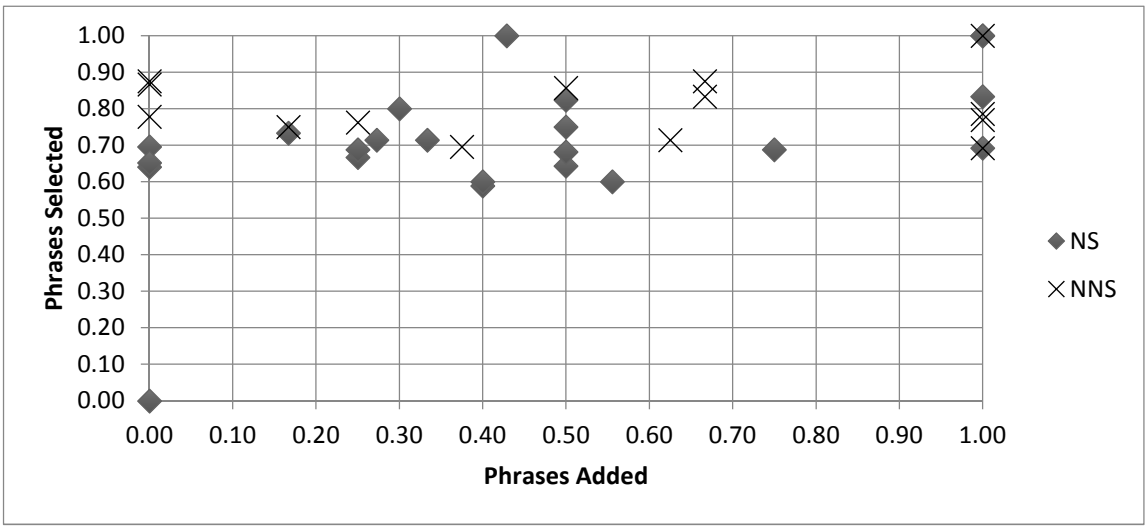

Figure 6: Comparison of precision in selected and added phrases with grammatical changes and abbreviations allowed as matches

Substitution of words, including the use of different prepositions, also results in sequences that were similar to those in the lecture, and still suggest that the sequences themselves were recalled, although not perfectly. Examples of sequences that show the word substitution characteristic are shown in Table 2. This category contains a total of 17 phrases, each of which is only added once by the participants. Figure 7 shows the changes in precision when phrases such as this are considered correct in addition to the grammatical changes. As expected, more of the participants have moved towards the right of the graph, but again, the participants have not reached the kind of precision figures seen for the task of selecting items from a provided list.

Phrase Added	**Closest Match in Transcript**
in murder cases	of murder cases
go from a to b	to go a to b
processional thinking	processual thinking
conventional idea	conventional view
this shift to	this shift towards

Table 2: Examples of phrases with similar word substitution

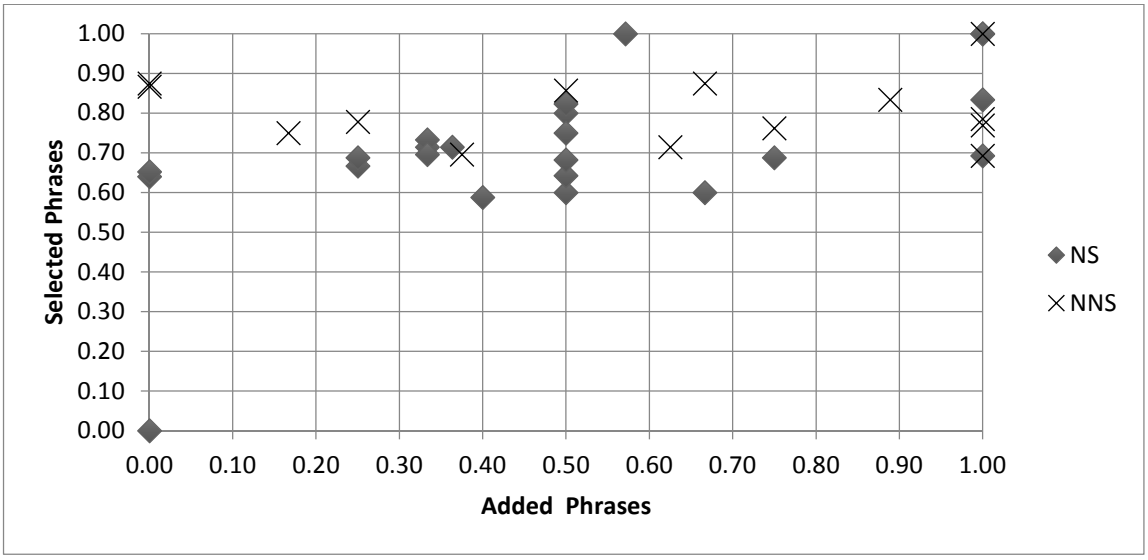

Figure 7: Comparison of precision in selected and added phrases with grammatical changes, abbreviations and small word changes allowed as matches

Overall, the participants were able to report a reasonable number of formulaic sequences beyond those supplied in the target list. However, this result has to be balanced by the difficulty they found in supplying the fully correct forms of the sequences they report. In many cases, it

seems that they remembered the *meaning* conveyed by the sequence, but could not accurately report the *form* of that sequence. Indeed, several of the participants made reference in the questionnaire to remembering and taking notes on concepts rather than exact language.

- 'I remember the concept not details.' *Non-native speaker*
- 'The way I approach lectures is not to remember isolated phrases unless prompted that it is particularly important but to take an overview.' *Native speaker*
- 'The thing that really challenges me here is that I was entirely focused on the content, preparing myself to answer questions about that, and so my notes were mainly phrases and quotes from the lecture rather than verbatim phrases. So I really have no idea what particular phrase he used.' *Native speaker*
- 'As I was listening as if to a normal lecture, it was the concepts and ideas to which I paid greater attention than the language.' *Native speaker*
- 'I don't [remember any phrases not on the list]! I remember a number of things being said but they were not exactly phrases, i.e. the way the list was.' *Native speaker*

Several core concepts from the lecture were picked up in the phrases added by the participants. These include:

- 'strategy being a pattern in a stream of past decisions'
- 'add a strategy after the event'
- 'implementation and strategy work in the same time'
- 'conventional, rational, orthodox way of defining'
- 'war on Iraq'
- 'key thinker'

With many phrases such as those given above, it is possible to see how they were triggered by the content of the lecture, but they did not actually occur as sequences in the transcript.

Discussion and Application

The motivation for this research was to investigate the potential of crowdsourcing for creating new types of context-specific spoken corpora, i.e. corpora made up of selected language sampled (in this case formulaic phrases) rather than full texts or episodes of discourse. We tested this with a focus on the ability to report formulaic sequences heard in an academic lecture. One of the important conclusions we can draw from this research is that while the individual seems to be better at selecting phrases heard from a list, the crowd is better at adding them. As was noted earlier, only 13 of the added phrases were added by more than one person. These phrases however do have a much higher chance of actually being heard in the text. If we take all of the phrases added and our most flexible approach to matching (grammatical and small word changes allowed), the percentage success is around 41%. However, if we take only those phrases added by more than one person, this success rate rises dramatically to around 85%. In addition, only two of these phrases were not reported in the exact form in which they appeared in the lecture, and even then only small grammatical changes were made. This supports the fundamental concept behind the use of crowd sourcing in this context, as by asking a large number of people, the results become more robust and reliable.

This insight suggests that a crowdsourcing approach to a contextually-sensitive phrase list may be a viable option. To deploy the study reported on in this paper in a real context of use would require further consideration of timing of the intervention via the smartphone. Informants could be prompted on a regular basis to note down the last phrase they had heard and/or to note down the next phrase that they will hear. This approach, while missing out on the richness of fully transcribed conversational data in context, affords an opportunity to develop a large-scale formulaic sequence list that is linked to geo-location and time. Importantly, the approach, as with other corpus-based methods, provides us with evidence of reported language usage rather than hypothetical accounts of usage and those based on intuition alone. As such it may also be used to replace a traditional Discourse Completion Task that focuses on formulaic phrases in context by moving from intuition to usage-based accounts. Further contextual variables could be added by the user so that a potential outcome could be a fully adaptive phrasebook that can be used in real time across different locations. A phrasebook such as this is likely to be appealing to language learners as they immerse themselves in a particular host culture using an app that allows them to contribute to the process of language documentation themselves. In sum, the approach outlined in this paper may help linguists develop more contextually sensitive descriptions of language in use by drawing on a new method of spoken language collection, as well as better applications based on those descriptions.

Notes

1. The research presented in this paper was supported by a research grant from the Engineering and Physical Science Research Council (EPSRC). Grant reference: EP/G065802/1; Grant title: Horizon: Digital Economy Hub at the University of Nottingham.
2. We are grateful to Anne Liu for her assistance in running the experiments.

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), pp 235–247.

Adda, G., Mariani, J. Besacier, L. and Gelas, H. 2014. *Crowdsourcing for speech: economic, legal and ethical analysis*. Open publication available online at: https://hal.archives-ouvertes.fr/hal-01067110.

Adda, G., Sagot, B., Fort, K. and Mariani, J-J. 2010. 'Crowdsourcing for Language Resource Development: Criticisms about Amazon Mechanical Turk overpowering use' in Proceedings of the Human Language Technology Challenges for Computer Science and Linguistics - 5th Language and Technology Conference (LTC 2011), pp 303–314. 25-27 November. Pozna, Poland.

Adolphs, S. and Carter, R. 2013. Spoken Corpus Linguistics: from monomodal to multimodal. London and New York: Routledge.

Ahn, L.V. 2006. 'Games with a purpose', *IEEE Computer Magazine* 36 (6), pp 92–94.

Aijmer, K. 2002. *English discourse particles: evidence from a corpus*. Amsterdam: John Benjamins.

Asheghi, N., Sharoff S. and Markert, K. 2014. 'Designing and Evaluating a Reliable Corpus of Web Genres via Crowd-Sourcing' in Proceedings of the Language Resources and Evaluation Conference (LREC 2014), pp 1339-1346. 26-31 May. Reykjavik: Iceland.

Callison-Burch, C. and M. Dredze. 2010. 'Creating speech and language data with Amazon's Mechanical Turk' in Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (NAACL HLT 2010), pp 1-12. 6 June. Los Angeles, California.

Callison-Burch., C. 2009. 'Fast, cheap and creative: Evaluating translation quality using Amazon's Mechanical Turk' in Proceedings of the Empirical Methods in Natural Language Processing (held in conjunction with ACL-IJCNLP) (EMNLP 2009), pp 286-295. 6-7 August. Suntec, Singapore.

Carter, R. A., and McCarthy, M. J. 1999. 'The English get passive in spoken discourse: description and implications for an interpersonal grammar', *English Language and Linguistics* 3 (1), pp 41-58.

Chamberlain, J., Poesio, M. and Kruschwitz, U. 2008. 'Phrase Detectives: a Web-based Collaborative Annotation Game' in Proceedings of the International Conference on Semantic Systems (I-Semantics 2008), pp 42-49. 3-8 September. Graz, Austria.

Evanini, K., D. Higgins, and K. Zechner. 2010. 'Using Amazon Mechanical Turk for transcription of Non-native speech' in Proceedings of the North American Chapter of the Association of Computational Linguistics Human Language Technologies Workshop (NAACL 2010), pp 53-56. 1-6 June. Los Angeles, California.

Gelas, H., Abate, S.T., Besacier, L. and Pellegrino, F. 2011. 'Evaluation of crowdsourcing for African languages' in Proceedings of the Conference on Human Language Technology for Development (HLT 2011), pp 128-133. 2-5 May. Alexandria, Egypt.

Goldman, J-P., Leemann, A., Kolly, M-J., Hove, I., Almajai, I., Dellwo, V. and Moran, S. 2014. 'A crowdsourcing smartphone application for Swiss German: putting language documentation in the hands of the users' in Proceedings of the Language Resources and Evaluation Conference (LREC 2014), pp 3444-3447. 26-31 May. Reykjavik, Iceland.

Gruenstein, A., I. McGraw, and A. Sutherland. 2009. 'A self-transcribing speech corpus: collecting continuous speech with an online educational game' in Proceedings of the ISCA Workshop on speech and language technology in education (SLaTE 2009). 3-5 September. Wroxall, UK.

Hughes, T., Nakajima, K., Ha, L., Vasu, A., Moreno, P. and LeBeau, M. 2010. 'Building transcribed speech corpora quickly and cheaply for many languages' in Proceedings of the Interspeech 2010, pp 1914–1917. 26-30 September. Chiba, Japan.

Ipeirotis, P. 2010. 'Demographics of mechanical turk'. *CeDER-10–01 working paper*. New York University.

Marge, M., Banerjee, S. and Rudnicky, A.I. 2010. 'Using the amazon mechanical turk for transcription of spoken language' in Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2010), pp 5270–5273. 14-19 March. Dallas, Texas.

Martinez, R. and Schmitt, N. 2012. 'A Phrasal Expressions List', *Applied Linguistics* 33 (3), pp 299-320.

McGraw, A. Gruenstein, A. and Sutherland, A. 2009. 'A self-labeling speech corpus: Collecting spoken words with an online educational game' in Proceedings of the Interspeech 2009, pp 3031-3034. 6-10 September. Brighton, UK.

McGraw, I., Lee, C., Hetherington, L., Seneff, S. and Glass, J. 2010. 'Collecting voices from the Cloud' in Proceedings of the Language Resources and Evaluation Conference (LREC 2010), pp 1576-1583. 17-23 May. Valletta, Malta.

Nattinger, J.R. and DeCarrico, J.S. 1992. *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Novotney, S. and Callison-Burch, C. 2010. 'Cheap, fast and good enough: automatic speech recognition with non-expert transcription' in Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp 207–215. 1-6 June. Los Angeles, California.

Pawley, A. and Syder, F.H. 1983. 'Two puzzles for linguistic theory: Nativelike selection and nativelike fluency' in J.C Richards and R.W. Schmidt (eds.) *Language and Communication*, pp. 191-225. London: Longman.

Raso, T. and Mello, H. (Eds.) 2014. *Spoken Corpora and Language Studies.* Amsterdam: John Benjamins.

Schauer, G. and Adolphs, S. 2006. 'Expressions of gratitude in corpus and DTC data: Vocabulary, formulaic sequences, and pedagogy', *System* 34(1), 119-134.

Schmitt, N. (ed.). 2004. *Formulaic Sequences*. Amsterdam: John Benjamins.

SDS *Sprachatlas der deutschen Schweiz*. 1962-2003. I-VIII. Basel: Francke.

Simpson-Vlach, R. and Ellis, N.C. 2010. 'An Academic Formulas List: New methods in phraseology research', *Applied Linguistics*, 31 (4), pp 487-512.

Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Appendix 1: Formulaic sequence list with frequency, frequency and category data

Phrase	selected by	NS	NNS	Content	target freq	BASE freq	1st	2nd	3rd
at the ___ level	4	1	3	Y	4	23	3	0	1
at the same time	3	0	3	N	1	96	0	0	1
back and forth	3	2	1	Y	2	3	0	0	2
corporate strategy	4	2	2	Y	2	2	2	0	0
decision making	4	1	3	Y	2	68	1	1	0
human beings	2	0	2	Y	2	59	2	0	0
in other words	3	0	3	N	1	349	0	0	1
ins and outs	1	0	1	N	1	0	0	1	0
key figure	4	2	2	N	1	3	0	0	1
long term	5	2	3	Y	4	68	0	0	4
made a mess	0	0	0	N	1	2	0	1	0
make sense	3	0	3	N	4	63	0	2	2
or whatever	2	0	2	N	2	193	1	1	0
over time	3	0	3	Y	5	47	0	0	5
pretty obvious	2	1	1	N	2	11	1	1	0
rather peculiar	0	0	0	N	1	1	0	0	1
so called	3	1	2	N	2	3	0	2	0
systems control	2	1	1	Y	3	0	3	0	0
top managers	5	2	3	Y	4	1	3	1	0
ups and downs	0	0	0	N	1	3	0	1	0
upside down	0	0	0	N	1	3	0	1	0
bear in mind	1	0	1	N	0	48			
first of all	3	0	3	N	0	214			
in a sense	2	0	2	N	0	128			
in practice	2	0	2	N	0	55			
in the sense that	2	0	2	N	0	37			
mind you	0	0	0	N	0	7			
on the other hand	2	0	2	N	0	87			
or something like that	0	0	0	N	0	62			
the extent to which	2	0	2	N	0	10			

xx = distractor items

Appendix 2: Pilot study questionnaire

Section A – About Yourself

1. Participant Number .....

2. Are you a native speaker of English? (please circle) Yes | No

2a. If no, what was your last IELTS/TEOFL score? .....

2b. When was this taken? .....

2c. What is your first language? .....

3. What course are you studying at Nottingham (include level i.e. BA, MA) .....

Section B – About the Task

4. How did you find the task you were asked to complete today? (please circle most appropriate answer)

Very easy *Easy* *Okay* *Difficult* *Very difficult*

5. To what extent were you confident that the phrases you selected were said during the lecture? (please circle most appropriate answer)

Fully *Partially* *Not at all*

6. Were you more confident about some phrases than others? (please circle) Yes | No

6a. If yes, can you remember any phrases you were particularly confident about?

Please list them below

6b. If yes, can you remember any phrases you were not very confident about?

Please list them below

7. Please explain in your own words how you approached the list of phrases. Did you for example read sequentially through the list and make a decision on each phrase at that time, read them all first and then make a decision, or use another strategy.

Section C – About the Software/Smart Phone

8. How easy to use did you find the software you used today? (please circle most appropriate answer)

Very easy *Easy* *Okay* *Difficult* *Very difficult*

9. Do you use a smart phone like this regularly? (please circle) Yes | No

10. Is there anything about the software that you feel could be improved?

Appendix 3: Main study questionnaire

Section A – About Yourself

1. Participant Number .....

2. Are you a native speaker of English? (please circle) Yes | No

2a. If no, what was your last IELTS/TEOFL score? .....

2b. When was this taken? .....

2c. What is your first language? .....

3. What course are you studying at Nottingham (include level i.e. BA, MA)

Section B – About the Task

4. How did you find the task you were asked to complete today? (please circle most appropriate answer)

Very easy *Easy* *Okay* *Difficult* *Very difficult*

5. To what extent were you confident that the phrases you selected were said during the lecture? (please circle most appropriate answer)

Fully *Partially* *Not at all*

6. Please explain in your own words how you approached the list of phrases. Did you for example read sequentially through the list and make a decision on each phrase at that time, read them all first and then make a decision, or use another strategy.

7. Below please list any phrases which you remember hearing in the lecture but that were not part of the list you were presented with on the phone.

Section C – About the Software/Smart Phone

8. How easy to use did you find the software you used today? (please circle most appropriate answer)

Very easy *Easy* *Okay* *Difficult* *Very difficult*

9. Do you use a smart phone like this regularly? (please circle) Yes | No

10. Is there anything about the software that you feel could be improved?