Journal of Risk Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rjrr20

# Using large text news archives for the analysis of climate change discourse: some methodological observations

Reiner Grundmann

**To cite this article:** Reiner Grundmann (2022) Using large text news archives for the analysis of climate change discourse: some methodological observations, Journal of Risk Research, 25:3, 395-406, DOI: 10.1080/13669877.2021.1894471

**To link to this article:** https://doi.org/10.1080/13669877.2021.1894471

JOURNAL OF RISK RESEARCH
2022, VOL. 25, NO. 3, 395–406
https://doi.org/10.1080/13669877.2021.1894471

Routledge
Taylor & Francis Group



# Using large text news archives for the analysis of climate change discourse: some methodological observations

Reiner Grundmann

Institute for Science and Society, School of Sociology and Social Policy, University of Nottingham, Nottingham, UK

**ABSTRACT**

This paper explores the contribution of software-based tools that are increasingly used for the semi-automated analysis of large volumes of text, especially Topic Modelling and Corpus Linguistics. These tools highlight the potential of getting interesting and new insights quickly, but at a cost. Linguistic aspects need to be considered carefully if computer-assisted technologies are to provide valid and reliable results. Main features of these tools will be presented, and some general problems and limitations will be discussed. The relation between technical tools and theoretical frameworks is discussed. The main empirical reference is the case of climate change.

**ARTICLE HISTORY**
Received 12 August 2020
Accepted 23 January 2021

**KEYWORDS**
Corpus linguistics; topic modeling; frame analysis; climate change discourses

## 1. Introduction

The aim of this paper is to explore the contribution of software-based tools that are increasingly used for the semi-automated analysis of large volumes of text.[1] I will provide an overview of two such technologies, Topic Modelling (TM) and Corpus Linguistics (CL). My main empirical reference is the case of climate change. While the literature review is not based on a systematic review it tries to present some of the most relevant contributions. I also include a more eclectic personal aspect to it which makes this a bit of an auto-ethnographic case study of a sociologist who became interested in linguistics and computer assisted tools. I will show that there is much to be gained by using such tools, but that they need to be aligned with theoretical perspectives and research questions in order to advance research in this field.

In this paper I will address two issues: one is methodological and deals with aspects of corpus construction, data analysis and significance of findings. The other focuses on two variables which could be considered crucial for research in this field, agency and frames. I will first of all summarize the central elements of TM and CL and then move on to sketch some common problems in large text data analysis, before moving on to the variables of interest.

## 2. Topic modelling

### *2.1. Scope and tools*

Topic modelling is a fast growing approach not only in computer science, but across the social sciences. Scholars use software packages to analyse large text databases for this task. Blei gives the following definition:

CONTACT Reiner Grundmann Reiner.Grundmann@nottingham.ac.uk Institute for Science and Society, School of Sociology and Social Policy, University of Nottingham, Nottingham, UK
This article has been republished with minor changes. These changes do not impact the academic content of the article.

'[M]achine learning researchers have developed probabilistic topic modeling, a suite of algorithms that aim to discover and annotate large archives of documents with thematic information. Topic modeling algorithms are statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time.' (Blei 2012, 77–78).

Maier et al. (2018, 93) summarize a major technique used in topic modeling, latent Dirichlet allocation (LDA). This 'is a computational content-analysis technique that can be used to investigate the "hidden" thematic structure of a given collection of texts. The data-driven and computational nature of LDA makes it attractive for communication research because it allows for quickly and efficiently deriving the thematic structure of large amounts of text documents. It combines an inductive approach with quantitative measurements, making it particularly suitable for exploratory and descriptive analyses.'

The tool is thus best described as 'semi-automated, unsupervised method for the analysis of textual content', in which 'the model attempts to "mimic" the writing process of a given corpus of documents. … "Topics" are statistical entities, sets of frequency distributions of words, based on the linguistic assumption of co-occurrence, that is, that words that are being used more frequently in the same documents also associate thematically. In topic models, every word in the corpus has a probability of appearing in each of the topics, and every document is composed as a mixture of all topics [ … ]. Topic modeling is a "bag-of-words" approach, which means that narrative, location in the text, and syntax are not taken into consideration.' (Walter and Ophir 2019, 254)

Several studies have applied LDA for the analysis of news reporting on climate change. Bohr (2020) created and analyzed a corpus of over 78,000 articles covering two decades from 52 US newspapers in order to investigate outlet bias with regard to geography, partisan orientation, or scale of circulation. Keller et al. (2020) used LDA on over 18,000 climate change articles published between 1997 and 2016 in two Indian newspapers. They categorized the news items into 28 different topics related to overarching themes such as impacts, science, politics, and society. They found that topics related to 'Climate Change and Society' and 'Climate Politics' appear more frequently than 'Climate Change Impacts' or 'Climate Science'. This raises the question what the differences between these headings are, and what they actually mean. The authors are aware of some limitations of LDA: The assignment of one prevalent topic to an article has to be considered with caution, as an article will rarely consist of only one topic. They also note that 'some topics are less robust, occur only in either one of the newspapers and are harder to validate by external reviewers which challenges their significance for this study' (Keller et al. 2020, 231). There are other studies available, such as Boussalis and Coan (2016) who present an analysis of US climate sceptical think tanks. Boussalis, Coan, and Poberezhskaya (2016) analze climate change coverage by the Russian press. Vu, Liu, and Tran (2019) provide a frame analysis of climate change reporting in 45 countries, based on LDA.

### 2.2. Problems and limitations

Topic modeling has received even more critical scrutiny by Brookes and McEnery (2019) who emphasize three main weaknesses: a lack of linguistic sensitivity, failure to define topics, and lack of replicability. 'In terms of linguistic sensitivity, the first feature worth considering is that topic modelling uses a very naïve model of a text. A text is viewed as being composed of a collection of words, a simple unordered set' (Brookes and McEnery 2019, 5). In a similar vein, Nerlich, Forsyth, and Clarke (2012, 48) observe the 'obvious fact that human language is inescapably a sequential phenomenon. Phonemes, words, phrases and other linguistic phenomena are generated and interpreted in a temporal sequence. A ''blind Venetian,'' for example, is not the same as a ''Venetian blind.''' A further problematic step is that LDA researchers often remove grammatical and function words, and use word lemmata, disregarding the different morphological expressions of a word. In doing so, important information can be lost. McEnery et al. (2015) show how the words muslim

and muslims 'consistently index different discourses about Muslims. Collapsing the frequency counts of both words together clearly risks at the very least blurring an important distinction and, at worst, losing or mischaracterising it.' In addition, some researchers remove words that are very frequent and very infrequent (see Maier et al. 2018).

Brookes and McEnery note the lack of any theoretical account of what constitutes a 'topic' within topic modelling: 'concepts like 'theme' and even 'coherence' are vague and remain ill-defined within this body of research. What constitutes a theme and what it is claimed makes that theme coherent are unclear' (Brookes and McEnery 2019, 6). And: 'in many existing topic model studies, topics are inferred by researchers without actually inspecting the texts assigned to that topic.' Murakami et al. (2017) also think that the term 'topic' is a misnomer, pointing out that '[t]hese groups of co-occurring words characterize "topics", and researchers may choose to refer to them using topic-like titles, but these are only convenient abstractions from lists of words' (Murakami et al., 2017, 244; see also Walter and Ophir 2019; Nicholls and Culpepper 2020).

The final criticism is that 'the replication – or more specifically, the repetition – of a topic model study is not necessarily possible. Consequently, a user may modify and re-run the topic modelling many times, evaluating outputs until finding an analysis that they deem to be credible and usable. This introduces the possibility of a high degree of subjectivity into the analysis.' (Brookes and McEnery 2019, 8). The problem of relicability has been addressed by others as well (see Roberts, Stewart, and Tingley 2016; Wilkerson and Casas 2017).

Their overall verdict is harsh: 'a major concern with topic modelling methods is the present lack of an adequate theoretical underpinning of what a topic actually is. This absence has given rise to ill-defined (and likely inconsistent) procedures of topic discovery that lack linguistic sensitivity and are, it seems, liable to make the false assumption that unordered and de-contextualised words can be mapped neatly onto propositional topics. Future research employing topic modelling methods should therefore endeavour to engage more deeply with linguistic theory when inferring the presence of topics in their textual data' (Brookes and McEnery 2019, 18).

DiMaggio, Nag, and Blei (2013) have applied this method claiming that it can help with frame analysis, and that it is well suited to analyzing polysemy and heteroglossia (polysemy refers to variations in meaning of a single term, heteroglossia refers to ambiguity at the level of the text). As the authors put it, a 'virtue of topic modeling is its deep affinity to the central insight in the sociology of culture that texts do not necessarily reflect a singular perspective but are often characterized by heteroglossia, the copresence of competing ''voices''—perspectives or styles of expression—within a single text' (DiMaggio, Nag, and Blei 2013, 582). In his seminal paper one of the co-authors of this paper had written that the fundamental 'intuition behind LDA is that documents exhibit multiple topics' (Blei (2012, 78). This indicates that some interesting work is going on in some areas of topic modeling.The reason may lie in the fact that the group of co-authors brings together a variety of expertise: 'Like any clustering technique, the method should be employed as a heuristic tool in combination with additional information by a research team that includes subject-area experts' (DiMaggio, Nag, and Blei 2013, 582). One might say that adding a linguist to the team might benefit the development of their methodology.

However, the techniques used for analysis need justification, as every methodology. The authors state that 'the sociology of culture has long been theory-rich and methods poor. Sociologists who study culture have generated numerous theoretical insights and developed concepts that promise a deep understanding of cultural change. Yet they have often lacked the means to make such concepts operational'—they suggest that LDA could provide the tools for the job (DiMaggio, Nag, and Blei 2013, 571). However, it is a large step from identifying a lack of operationalization of theoretical concepts that could be used for empirical research to the assertion that one specific software application would provide the answer.

What is more, the empirical case used in their paper, the news coverage of public funding of the arts in the US, is underwhelming as regards the results. The main finding is that the news coverage focused on controversy, 'producing a cloud of negative representations', especially after

the election of George H.W. Bush, which unleashed a 'culture war'. The question is how unique this insight is, and if it goes beyond what an informed news reader would already know.

The authors recognize this problem and state in concluding that 'the model is just the beginning. For cultural analysis, the purpose of modeling is to apprehend the structure of the data and render it tractable by producing meaningful topics … that can be used to answer more focused questions … Topic modeling will not be a panacea for sociologists of culture. But it is a powerful tool for helping us understand and explore large archives of texts' (DiMaggio, Nag, and Blei 2013, 602–3). What they say about cultural sociology applies to other social science fields which are theory rich but methods poor. Topic modelling is unlikely to provide a ready-made tool to fill this gap. Operationalization of key theoretical concepts remains a major task.

## 3. Corpus linguistics

### *3.1. scope and tools*

There is no agreed definition of what CL is or does (see Gries 2009). I use the definition of McEnery and Andrew (2012, 1–2) who state:

'We could reasonably define corpus linguistics as dealing with some set of machine-readable texts which is deemed an appropriate basis on which to study a specific set of research questions. The set of texts or corpus dealt with is usually of a size which defies analysis by hand and eye alone within any reasonable timeframe. It is the large scale of the data used that explains the use of machine-readable text … corpora are invariably exploited using tools which allow users to search through them rapidly and reliably. Some of these tools, namely concordancers, allow users to look at words in context. Most such tools also allow the production of frequency data of some description.'

Word frequencies, word collocations and clusters, and key words are prominent metrics in CL research. Word frequencies indicate the importance of specific words within a given text. Collocations indicate how a specific word is accompanied by other words–its collocates– within close proximity in a text. Word clusters are (short) strings of words occurring in a text, also called n-grams. Key word analysis establishes the relative salience of specific words from one selected text compared to a reference corpus.

There is a limited number of studies that have used CL methods for the analysis of climate change discourse. Brigitte Nerlich and a team of co-authors have investigated English speaking news coverage, looking at linguistic properties, metaphors, and readers' comments. (Collins and Nerlich 2015; Nerlich and Koteyko 2009; Koteyko, Jaspal, and Nerlich 2013; Jaspal, Nerlich, and Koteyko 2013). Dayrell (2019) built a corpus of Brazilian news reports on climate change consisting of 19,686 newspaper texts (11.4 million words) published by 12 Brazilian broadsheet papers. She used keyword and collocation analysis for the investigation. Salient keywords are greenhouse gas emissions, UN conferences, fossil fuels, renewable energy or deforestation.

### *3.2. Problems and limitations*

Published CL work uses corpora built from downloads of news archives, typically from newspapers or aggregator sites like Nexis. Availability of full-text material allows the analysis of large amounts of material, without the need for sampling. The downside is that contextual information is lost, or not immediately available, such as pictures, text boxes, or graphical presentation. Another problem is the 'noise' contained in the downloads, and the question how much noise should be eliminated in the final ('cleaned') corpus. The answer will depend on how one answers two other questions: Should the corpus only contain relevant stories? And: What is relevant?

Most research is not concerned with cleaning and vetting the downloaded items, apart from eliminating duplicates. Nerlich, Forsyth, and Clarke (2012, 47) state: 'We decided not to clean our corpus by filtering out texts containing […] apparently irrelevant phrases […], to avoid biasing

our findings by what we expected to find. In this respect our philosophy is, as far as possible, to ''let the data speak for themselves'' even if it means accepting a certain level of noise in the corpus.' I will return to this issue in the next section.

Another problem is the identification of meaning and voices within a corpus. Research on CC discourse frequently presents a 'pro' or 'anti' stance in news. However, it is not easy to extract this kind of information from an article. Often several messages are conveyed, and it can be difficult to ascertain what the journalist's voice and position is.

As Dahl and Kjersti (2014) argue, in any given text there is linguistic polyphony. 'The issue is approached from different perspectives by social actors with different backgrounds, world views, interests, values, and beliefs (Hulme 2009), a situation which implies position taking. Thus, both voices and positions become important objects of study in order to understand the complexity of the debate' (Dahl and Kjersti 2014, 402).

They explain the idea of linguistic polyphony 'that in one single utterance there may be several voices or points of view present, in addition to the one of the speaker/writer' (Dahl and Kjersti 2014, 402).

There are other linguistic properties which pose a difficulty for textual analysis. One of them is the use of irony, satire or humour. Taken together, this means that words cannot be taken at face value and their meaning can only be grasped in context (Partington 2007).

Nevertheless, researchers have tried to overcome such doubts and develop methods for linguistic analysis of news texts, partly with the help of software systems. Semantic tagging is a method which categorizes words into a classificatory scheme. It has been used by researchers, following seminal work at Lancaster University (http://ucrel.lancs.ac.uk/usas/).

Collins and Nerlich (2015) analysed reader comments on climate change in the Guardian, using semantic tagger software WMatrix. This software tool compares words with key categories derived from British National Corpus. The authors praise the advantage of this process, being 'systematic and automatic, organizing thousands of words of data into semantic categories in a matter of seconds' (Collins and Nerlich 2015, 195). However, this approach potentially underestimates the challenge of ambiguity of meaning.

## 4. Common issues and problems

In this section I want to draw on my own experience with one of the methods (CL), identify two methodological issues, and suggest some steps for further research. Some parts of this section are more relevant to CL, but there are links to TM as well. The next section (5) will then look at the problem of agency and visibility, and the section thereafter (6) at the issue of identifying frames.

### *4.1. Corpus construction*

In previous work we wanted to establish sound methodological rules for downloading and preparing corpora for analysis (Grundmann and Scott 2014; Grundmann, Kreischer, and Scott 2016). Our aim was to construct a reliable dataset of news items that excludes bias as far as possible, including as many items as possible on a given issue, and only on the issue. Our dataset is based on news archives of written text (sometimes only on one type of newspaper, such as prestige press, as is the case with our work on austerity), and thus excludes other sources, which might merit analysis such as broadcasts, visuals, or social media comments. In this sense our data is biased towards one type of source. Nevertheless, print and prestige press stories are important because political elites tend to read them, and journalists report about political elites and provide reports and quotes from institutional actors.

However, the construction of the dataset follows a procedure which does not rely on subjective choice. It also tries to include as many sources as possible, and not restrict the volume

through random or purposive sampling, through lemmatization, or through exclusion of frequent or infrequent words (see also Denny and Spirling 2018).

After gaining experience with the method and after the publication of our first paper (Grundmann and Scott 2014) we posed ourselves the question: How should a textual corpus be constructed? How much noise should be tolerated? Our approach was guided by the principle to include only texts which are really news stories about the subject. We developed a parser which performs this task automatically. We applied this method to the construction of a corpus on the topic if austerity in the British press and are repeating this procedure for CC. This is a time-consuming process, but we think it is necessary. How can we be sure we are analysing CC discourse if a large proportion of a downloaded data consists of irrelevant news items? Our estimate is that the proportion of real climate stories is much less than 60% of a standard Nexis download based on a keyword search including 'climate change' or 'global warming'.[2]

The above cited work by Bohr (2020) and Keller et al. (2020) is cognizant of this problem. Bohr describes how he first excluded duplicates from an initial download based on a Boolean search for 'climate change' or 'global warming'. After manual inspection it became clear that 'many articles did not focus on climate change and simply mention one of the keyword phrases a single time in passing. An additional filtering procedure required an article to mention the terms 'climate change,' 'global warming,' or 'greenhouse gas' at least twice. This reduced the corpus to less than 50% of the original download.

Keller et al. (2020, 231) state about their work that articles only mentioning climate change in passing could have been included in their sample. As a remedy they suggest 'using sharper identification criteria, for example more specific search terms' for future studies.

### *4.2. Key word analysis*

The use of an appropriate reference corpus is important for the calculation of key words. Many researchers use a standard reference corpus, such as the British National Corpus. In the case of climate change, climate relevant words will be salient, such as climate, climate change, global warming, IPCC, temperatures, conferences, Kyoto, and so on. Such terms are to be expected and the analysis merely confirms what every news reader will already know.

In order to identify more specific key words we decided to use the whole climate corpus (our downloaded and cleaned dataset) as the reference corpus. The resulting keyword analysis provides results that go beyond the obvious as they show salient words from within a subset of the database. Using this procedure, obvious words and word combinations like 'climate change' or 'global warming' are no longer salient, but words such as 'developing', 'trading', 'subsidies', or 'transport'.

## 5. Actors and claims makers

I got involved with CL scholarship after publishing an article in Environmental Politics (Grundmann 2007). In this piece I was interested in the visibility of specific actors in the CC discourse in Germany and USA, namely advocates, sceptics and the IPCC. This research interest was informed by a pre-theoretical, implicit assumption that the public visibility of sceptical voices about climate change would hamper the development of climate policies. Such a belief was, and still is, widespread among social science researchers, climate scientists, and climate activists. The comparison between the US, and Germany seemed to show that the different levels of public visible scepticism correlated with the ambition of climate policies in these two countries. An influential article on US news coverage of climate change had shown that the visibility of sceptics was artificially inflated through journalistic practices of providing false balance in news reporting (Boykoff and Boykoff 2004). As it turns out, the higher visibility of sceptics in the US reflects its political constitution, where robust and open (some might say: adversarial) debates about the scientific evidence for policy is common practice. This means that the causal

relationship between visibility and efficacy needs to be examined more carefully than I had assumed. The second problem with this approach is empirical: even in the US it turned out that the postulated journalistic norm of 'false balance' was not in evidence after 2005 (Boykoff 2007).

This raises the question of agency and visibility in a given corpus. In order to identify relevant and potentially dominant actors we initially used lists of specific claims makers which we drew up in advance (Grundmann and Scott 2014). This introduces a potentially large dose of arbitrariness which ideally should be avoided. In order to do so, we decided to use Named Entity Recognition (NER) software to extract names of claims makers from our dataset. We applied this in research about the discourse on Austerity in the British Press (Grundmann, Kreischer, and Scott 2016) and are using it in an ongoing project on CC.

It is worth mentioning that this new procedure is inductive and does not rely on inferring topics or meaning to items found in our corpus. Relevant social actors (individuals and organizations) are identifiable directly.

In our research on the discourse on austerity in the British press we also analyzed verb collocations of central actors (Grundmann, Kreischer, and Scott 2016). We found, for example, that the Chancellor of the Exchequer, George Osborne played a dominant role. He had by far the most mentions in the news corpus: he was quoted ten times more frequently than opposition speakers such as Ed Miliband or Gordon Brown. Not only the word frequency testifies to this role, it is also indicated by the word collocates 'announced', 'unveiled', 'delivers', 'prepares', 'introduces', and 'urged'. In contrast, collocates for Gordon Brown were 'used', 'introduced', 'was', and 'tried'. This shows a difference in agency and ambition, as seen through the press coverage of their statements in public.

Our work on the austerity discourse also shows the need to be cautious with regard to the visibility and agency of claims makers. As the example above suggests, visibility does not equate agency, or more precisely: political influence. The most visible claims makers in our dataset on austerity were the Chancellor of the Exchequer and the Labour party. While the former had agency in defining situations, coining metaphors, phrases, and narratives, the latter did not. As our data shows, it was absorbed by its internal leadership struggles and did not come up with a narrative to challenge the Chancellor.

The high visibility of actors does not mean they are effective. They may be visible in public discourse but such visibility need not translate into political influence, or power. A contextual analysis of their public appearance is necessary. This can be done, to some extent, through collocation analysis. There might also be institutional or administrative power holders behind the scenes who do not manifest themselves in public discourse. In addition, these methods do not allow an analysis of layers of meaning and positioning, as shown by Fløttum and Dahl. Visibility of climate sceptics, to use an overused example, does not mean that sceptical voices are celebrated. They could be mentioned because they are strongly criticised, thus giving them salience. The upshot of this is that we need more empirically informed analysis in order to address theoretical issues about agency, visibility, and power (see the three dimensions of power identified by Lukes 2004). Theories of power and of the policy process could benefit greatly from such work, especially if it is comparative in nature.

Our research shows different salient words across countries (Grundmann and Scott 2014). We also observe different claims makers, even different types of claims makers dominating public discourse in the countries we analysed. In Germany and the UK environmental NGOs are dominant, in the US it is politicians and celebrities (Al Gore, Arnold Schwarzenegger). In France international organizations, such as the IPCC, get a lot of mentions. These findings can be interpreted on the basis of existing theoretical frameworks, such as civic epistemologies. The term indicates the 'institutionalized practices' through which actors, or claims makers 'test and deploy knowledge claims used as a basis for making collective choices' (Jasanoff 2005, 255). Those institutionalized practices differ from country to country.

Visible claims makers are often limited to their domestic audience, with few exceptions. In our four-country comparison the USA is the only country whose claims makers are mentioned in the other three countries. The reverse is not true. This shows the central role the US plays in

global climate science and politics. These results could be brought into productive dialogue with the concept of 'domestication' of climate politics (Olausson 2014).

During some time periods we observe that claims makers and/or topics are converging; we call this the synchronization of transnational discourse (Grundmann, Smith, and Wright 2000). We found for example, that in 2005, 2008, and 2009 there was a convergence of key words across the four countries in our sample. The same is not true for other periods. Here, the discourse is shaped by other events, and visible through different key words in different countries.

This indicates that climate change is embedded in national discourses in specific ways. In our data climate change is 'riding' on topics unique to each country. A reasonable hypothesis is that the discourse is shaped differently in different countries in 'normal times', and only comes together in exceptional periods. Hence, we should see discursive divergence over several years. Our previous analysis points in this direction, although the time period is relatively short (2005–2010). These observations show how only at critical junctures the discourses across nations become synchronized. In most years the climate discourses are nationally specific. This observation should be tested in future work, over longer periods time, using shorter intervals (monthly data).

## 6. Frames

Frame analysis is a central element of much work in the field of Discourse analysis. However, frames are typically constructed manually, and no software exists which could extract frames from texts (Nicholls and Culpepper 2020). Although DiMaggio et al. have tried to derive frames from topics, commentators are sceptical about their claims (see Walter and Ophir 2019; Maier et al. 2018 for a critical evaluation).

The challenge is illustrated by a publication that performs manual frame identification. In their analysis of the Fifth Assessment Report published by the IPCC in 2014, O'Neill et al. (2015) identify, based on previous studies, ten frames: Settled Science (SS), Political or Ideological Struggle (PIS); Role of Science (ROS); Uncertain Science (US); Disaster (D); Security (S); Morality and Ethics (ME); Opportunity (O); Economic (E); and Health (H). These frames were operationalized so that coders could identify them, using specific criteria. However, as often is the case with coding schemes, these are not mutually exclusive. For example, the coding item 'Unprecedented rate of change compared to palaeo records' appears under SS, and 'Unprecedented rise in global average surface temperature' appears in D. The demarcation between the two could be non-trivial. Likewise, 'questioning the motives or funding of opponents' (coded as PIS) could be similar to 'Urges trust in climate scientists and dismisses sceptic voices' (coded as SS).

In theory, coding procedures could be refined to avoid ambiguity, transferred into a software algorithm, and via machine learning processes be perfected to a stage where results are (re-) producible in unsupervised ways. As mentioned above, some researchers using LDA maintain that these technical developments may become reality soon, but do not exist at present.

We suggest a different approach. Following the literature on the construction of social problems (Spector and Kitsuse 2001; Trumbo 1996) we are interested in those claims makers who are prominent in public discourse. Ideally, we would like to know who the claims makers are, and what claims they make. While appropriate software such as NER allows us to identify persons and organizations, the identification of claims from the texts is not possible with software, at least not in a direct way. We can get close by inspecting word collocations but claims can only be inferred from this information. We are facing the problem of inferring meaning from word patterns, establishing rules and algorithms where possible, by searching the database.

Matthes and Kohring (2008) present an approach which manually identifies frame elements, which are parts of a bigger frame. These are based on Entman's definition of frames as a compound of 'problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described.' (Entman 1993, 52). They then apply cluster analysis to the frame elements and construct frames in this way. They convincingly argue that identifying frame

elements leads to more reliable results compared to a holistic frame analysis. They state: 'The crucial difference to the common assessment of frames is that frames are empirically determined and not subjectively defined.' (Matthes and Kohring 2008, 265). They identify topics and actors, and operationalize the four frame elements through a coding procedure.

However, their procedure is labour intensive and limited to a sample that can be researched in a given time by human researchers. It is time consuming and costly to identify topics and actors, and to operationalize the four elements of Entman's definition. They also use the article as unit of analysis which can be problematic as many news stories have a range of actors and/or topics. The critical comments above (see Dahl and Fløttum 2014) apply. Some researchers using LDA recognize the problem, too (DiMaggio, Nag, and Blei 2013).

Consider the following statement: 'Climate change is human-made and getting worse, we need urgent action.' This statement, or variations of it, can be found in many news stories, sometimes attributed to one actor, sometimes reflecting the gist of the article. The statement combines causal attribution and moral evaluation. Statements like these pose coding problems as the message to be extracted falls under two frame elements. What is more, the time dimension is missing from the frame elements. After all, many actors would agree that climate change is a real, and serious issue, but could disagree about how urgent action is, or what costs would be justified.

Another typical statement, 'Fossil fuels are to blame for the climate crisis, we need to ramp up our efforts and expand renewable energy' combines problem definition, causal interpretation, moral evaluation and treatment recommendation. In this version the distinction between frame elements becomes blurred, leading to ambiguity in coding.

The upshot of this discussion is that Entman's conceptual distinctions of frames has been useful for informing empirically informed frame analysis, but also has clear limitations. It cannot be regarded as a universal tool as it neglects the time dimension and is under-complex with regard to the treatment options. In the case of climate change there is political and scientific argument about the merit of specific solutions, with no consensus emerging. This is not surprising if one considers climate change as a 'wicked problem' which cannot be solved, but only managed better or worse (Newman and Head 2017; Hoppe, Wesselink, and Cairns 2013; Rayner 2006; Grundmann 2016).

This would call for a more refined definition of frame elements, paying special attention to the *treatment* aspect. Here several policy options have emerged; I have identified a dozen: 1) rolling out nuclear power plants across the globe; 2) switching all energy supply to solar, wind, or biofuels; 3) taxing carbon (or energy) with a) low or b) high rates; 4) implementing emission-trading systems; 5) developing carbon-capture and storage technologies; 6) developing new zero-carbon energy systems; 7) taking adaptation more seriously; 8) developing geo-engineering projects; 9) adopting vegetarian or vegan diets and lifestyles; 10) restricting population growth; 11) abolishing capitalism; 12) abolishing democracy (Grundmann 2018).

Coding news items into this, or any other conceptual scheme, is a deductive approach which crucially rests on coding procedures that require robust agreement between coders about the meaning of an item. CL is not able to fulfil this task. But it can offer a different methodological option, based on its unique features: being inductive, either corpus-driven or corpus based (Tognini-Bonelli 2001).

This is facilitated by the availability of a list of claims makers, obtained through NER parsing. This then can be complemented by collocation analysis, showing collocations for the highly visible claims makers. While this approach does not allow to identify frames in unsupervised settings, it shows options to identify frame elements in a more detailed way.

## 7. Conclusion

If we look back at the points made so far, it appears that there are some common problems, some problems unique to each approach, and some tentative suggestions about a way forward.

There are several advantages to use computer assisted methods. They allow to map a discourse over time, providing a 'helicopter' perspective. Major features of texts can be detected

through software algorithms, not relying (too much) on subjective decisions of researchers. The analysis aims to discover hidden structures, either through topic analysis or patterns of words which are not visible to the reader. As John Sinclair put it memorably, 'Language users cannot accurately report language usage, even their own' (Sinclair 1994, cited in Willis 2017). Results can be represented in a summary way, in timelines, and comparisons between countries or types of media outlets are within easy reach of researchers.

A disadvantage using computational methods is the fact that one does not necessarily get beyond the surface, or that trivial findings are obtained. If one wants to avoid these pitfalls, several methodological decisions have to be made. These have to do with the construction of the corpus: which texts should be included, from which sources, over what time period? Statistical measures need to be identified and justified. The reduction of an enormous amount of information to a manageable size, i.e. a size that can be presented in a meaningful way and interpreted is a challenge that requires judgement. And the interpretation of results is still necessary; no text exists without context. In many cases a comprehensive database will provide contextual information as single articles can be inspected for closer analysis. This is possible if a full text database is available throughout the research process. A combination of quantitative and qualitative work seems therefore appropriate.

The lure of computer assisted analysis of vast amounts of texts has led some researchers to explore software applications which are often black boxes for the user. The appeal to objectivity via automation is limited by theoretical assumptions such as the 'bag-of- words-approach' (which is also present in CL keyword methodology, see van Meter 2018), or the exclusion of word variations and function words, or very frequent and infrequent words as 'noise'. The use of linguistic software without linguistic expertise is problematic, yet it seems appealing as it allows to produce quick results.

Maybe the numerical packages, from Excel, SPSS to R can teach a lesson. They provide different levels of data analysis and visualization, and users are still able to make the best, or worst of it. A quick and dirty statistical analysis will not withstand the scrutiny of a large and competent expert community. Suggesting causation by calculating correlations is a commonly exposed flaw. Language based software is no different. It provides tools for analysis and presentation which vary in quality. To check the quality of this research we need a critical mass of competent practitioners . We need in-depth, interdisciplinary efforts between linguists, programmers, and social scientists in order to enhance the research potential of these new and exciting technical opportunities.

Practitioners need to ask themselves what kind of questions they want to examine. As the example of topic modelling indicates, there is a danger that the availability of technical applications leads to research questions that are somehow new, but produces underwhelming results. Discourse researchers have been keen to investigate the narratives and meanings of texts, how issues are framed by actors, how types of communication are used in efforts of persuasion, and what these results actually mean for social reality.

Computer assisted technologies for textual analysis should fulfil two tasks: they should help us do a better job with research questions we have always been interested in. And they should provide us with new possibilities, and unexpected insights. Topic modelling, as the name suggests, is interested in the structural properties of texts, not specifically in the efforts of persuasion by specific social actors. CL can offer a powerful tool to analyse large amounts of text. It is very good at providing quick overviews and pointers which one may not have considered. However, it needs to be complemented by careful study—datasets need to be cleaned, lists of named entities need to be vetted and amalgamated, methods for the choice of keywords need to be established, and meaning needs to be constructed on the basis of close inspection. All methods have to be applied and guided by justified decisions for a particular purpose. This means, these decisions have to be replicable, accessible and clearly expressed. Research teams need to be built which have the requisite variety of expertise and are able to address the problems and limitations identified in this paper.

## Notes

1. I would like to thank Mike Scott, an anonymous reviewer, and the editor for helpful suggestions.
2. Readers who are interested in exploring the tools we have used should look here https://lexically.net/wordsmith/ for access to the program. For the processing:
   parsing Nexis & Factiva: https://lexically.net/DownloadParser/index.htm
   finding duplicates: https://lexically.net/downloads/version8/HTML/different_contents_dup_finder.html
   checking content: https://lexically.net/downloads/version8/HTML/relevance_check.html
   building monthly or yearly sub-corpora: https://lexically.net/downloads/version8/HTML/build_sub_corpora.html
   key words and time-lines: https://lexically.net/downloads/version8/HTML/kwdb_database_timelines.html

## ORCID

Reiner Grundmann http://orcid.org/0000-0003-0266-9296

## References

Blei, David. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77–84. doi:10.1145/2133806.2133826.

Bohr, Jeremiah. 2020. "Reporting on Climate Change: A Computational Analysis of U.S. Newspapers and Sources of Bias, 1997–2017." *Global Environmental Change* 61 (December 2019): 102038. doi:10.1016/j.gloenvcha.2020.102038.

Boussalis, Constantine, and Travis G. Coan. 2016. "Text-Mining the Signals of Climate Change Doubt." *Global Environmental Change* 36: 89–100. doi:10.1016/j.gloenvcha.2015.12.001.

Boussalis, Constantine, Travis G. Coan, and Marianna Poberezhskaya. 2016. "Measuring and Modeling Russian Newspaper Coverage of Climate Change." *Global Environmental Change* 41: 99–110. doi:10.1016/j.gloenvcha.2016.09.004.

Boykoff, Maxwell. 2007. "Flogging a Dead Norm? Newspaper Coverage of Anthropogenic Climate Change in the United States and United Kingdom from 2003 to 2006." *Area* 39 (4): 470–481. doi:10.1111/j.1475-4762.2007.00769.x.

Boykoff, Maxwell, and Jules Boykoff. 2004. "Balance as Bias: Global Warming and the US Prestige Press." *Global Environmental Change* 14 (2): 125–136. doi:10.1016/j.gloenvcha.2003.10.001.

Brookes, Gavin, and Tony McEnery. 2019. "The Utility of Topic Modelling for Discourse Studies: A Critical Evaluation." *Discourse Studies* 21 (1): 3–21. doi:10.1177/1461445618814032.

Collins, Luke, and Brigitte Nerlich. 2015. "Examining User Comments for Deliberative Democracy: A Corpus-Driven Analysis of the Climate Change Debate Online." *Environmental Communication* 9 (2): 189–207. doi:10.1080/17524032.2014.981560.

Dahl, Trine, and Fløttum Kjersti. 2014. "A Linguistic Framework for Studying Voices and Positions in the Climate Debate." *Text and Talk* 34 (4): 401–420. doi:10.1515/text-2014-0009.

Dayrell, Carmen. 2019. "Discourses around Climate Change in Brazilian Newspapers: 2003–2013." *Discourse & Communication* 13 (2): 149–171. doi:10.1177/1750481318817620.

Denny, Matthew J., and Arthur Spirling. 2018. "Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It." *Political Analysis* 26 (2): 168–189. doi:10.1017/pan.2017.44.

DiMaggio, Paul, Manish Nag, and David Blei. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41 (6): 570–606. doi:10.1016/j.poetic.2013.08.004.

Entman, R. M. 1993. "Framing: Towards Clarification of a Fractured Paradigm." *Journal of Communication* 43 (4): 51–58. doi:10.1111/j.1460-2466.1993.tb01304.x.

Gries, Stefan Th. 2009. "What is Corpus Linguistics?" *Language and Linguistics Compass* 3 (5): 1225–1241. doi:10.1111/j.1749-818X.2009.00149.x.

Grundmann, Reiner. 2007. "Climate Change and Knowledge Politics." *Environmental Politics* 16 (3): 414–432. doi:10.1080/09644010701251656.

Grundmann, Reiner. 2016. "Climate Change as a Wicked Social Problem." *Nature Geoscience* 9 (8): 562–563. doi:10.1038/ngeo2780.

Grundmann, Reiner. 2018. "Ozone and Climate Governance: An Implausible Path Dependence." *Comptes Rendus Geoscience* 350 (7): 435–441. doi:10.1016/j.crte.2018.07.008.

Grundmann, Reiner, Kim-Sue Kreischer, and Mike Scott. 2016. "The Discourse of Austerity in the British Press." *Zeitschrift Für Politik*. Special Issue 8: 92–127.

Grundmann, Reiner, and Mike Scott. 2014. "Disputed Climate Science in the Media: Do Countries Matter?" *Public Understanding of Science (Bristol, England)* 23 (2): 220–235. doi:10.1177/0963662512467732.

Grundmann, Reiner, Dennis Smith, and Sue Wright. 2000. "National Elites and Transnational Discourses in the Balkan War." *European Journal of Communication* 15 (3): 299–320. doi:10.1177/0267323100015003003.

Hoppe, Rob, Anna Wesselink, and Rose Cairns. 2013. "Lost in the Problem: The Role of Boundary Organisations in the Governance of Climate Change." *Wiley Interdisciplinary Reviews: Climate Change* 4 (4): 283–300. doi:10.1002/wcc.225.

Jasanoff, Sheila. 2005. *Designs on Nature: Science and Democracy in Europe and the United States*. Princeton, N.J.: Princeton University Press.

Jaspal, R., B. Nerlich, and N. Koteyko. 2013. "Contesting Science by Appealing to Its Norms: Readers Discuss Climate Science in the Daily Mail." *Science Communication* 35 (3): 383–410. doi:10.1177/1075547012459274.

Keller, Tobias R., Valerie Hase, Jagadish Thaker, Daniela Mahl, and Mike S. Schäfer. 2020. "News Media Coverage of Climate Change in India 1997–2016: Using Automated Content Analysis to Assess Themes and Topics." *Environmental Communication* 14 (2): 219–235. doi:10.1080/17524032.2019.1643383.

Koteyko, Nelya, Rusi Jaspal, and Brigitte Nerlich. 2013. "Climate Change and 'Climategate' in Online Reader Comments: A Mixed Methods Study." *The Geographical Journal* 179 (1): 74–86. doi:10.1111/j.1475-4959.2012.00479.x.

Lukes, Steven. 2004. *Power: A Radical View*. Second ed. London: MacMillan.

Maier, Daniel, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, et al. 2018. "Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology." *Communication Methods and Measures* 12 (2-3): 93–118. doi:10.1080/19312458.2018.1430754.

Matthes, Joerg, and Matthias Kohring. 2008. "The Content Analysis of Media Frames: Toward Improving Reliability and Validity." *Journal of Communication* 58 (2): 258–279. doi:10.1111/j.1460-2466.2008.00384.x.

McEnery, Tony, and Hardie Andrew. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Meter, Karl M. van. 2018. *Obama Leads to Trump: 2015 - 2017 World Media Analysis*. Paris: Editions L'Harmattan.

Murakami, Akira, Paul Thompson, Susan Hunston, and Dominik Vajn. 2017. "What is This Corpus about?': Using Topic Modelling to Explore a Specialised Corpus." *Corpora* 12 (2): 243–277. doi:10.3366/cor.2017.0118.

Nerlich, Brigitte, Richard Forsyth, and David Clarke. 2012. "Climate in the News: How Differences in Media Discourse between the US and UK Reflect National Priorities." *Environmental Communication* 6 (1): 44–63. doi:10.1080/17524032.2011.644633.

Nerlich, Brigitte, and Nelya Koteyko. 2009. "Compounds, Creativity and Complexity in Climate Change Communication: The Case of 'Carbon Indulgences." *Global Environmental Change* 19 (3): 345–353. doi:10.1016/j.gloenvcha.2009.03.001.

Newman, Joshua, and Brian W. Head. 2017. "Wicked Tendencies in Policy Problems: Rethinking the Distinction between Social and Technical Problems." *Policy and Society* 36 (3): 414–429. doi:10.1080/14494035.2017.1361635.

Nicholls, Tom, and Pepper D. Culpepper. 2020. "Computational Identification of Media Frames: Strengths, Weaknesses, and Opportunities." *Political Communication* 00 (00): 1–23. doi:10.1080/10584609.2020.1812777.

O'Neill, Saffron, Hywel T. P. Williams, Tim Kurz, Bouke Wiersma, and Maxwell Boykoff. 2015. "Dominant Frames in Legacy and Social Media Coverage of the IPCC Fifth Assessment Report." *Nature Climate Change* 5 (4): 380–385. doi:10.1038/nclimate2535.

Olausson, Ulrika. 2014. "The Diversified Nature of 'Domesticated' News Discourse: The Case of Climate Change in National News Media." *Journalism Studies* 15 (6): 711–725. doi:10.1080/1461670X.2013.837253.

Partington, Alan. 2007. "Irony and Reversal of Evaluation." *Journal of Pragmatics* 39 (9): 1547–1569. doi:10.1016/j.pragma.2007.04.009.

Rayner, Steve. 2006. "Wicked Problems: Clumsy Solutions- Diagnoses and Prescriptions for Environmental Ills." *Jack Beale Memorial Lecture on Global Environment* 1–12. https://core.ac.uk/download/pdf/288283455.pdf.

Roberts, Margaret E., Brandon M., Stewart, and Dustin Tingley. 2016. "Navigating the Local Modes of Big Data: The Case of Topic Models." In *Computational Social Science*, edited by R. Michael Alvarez, 51–97. New York, NY: Cambridge University Press. doi:10.1017/cbo9781316257340.004.

Spector, Malcolm , and John I. Kitsuse. 2001. *Constructing Social Problems*. London: Cummings.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: Benjamins.

Trumbo, Craig. 1996. "Constructing Climate Change: Claims and Frames in US News Coverage of an Environmental Issue." *Public Understanding of Science* 5 (3): 269–283. doi:10.1088/0963-6625/5/3/006.

Vu, Hong Tien, Yuchen Liu, and Duc Vinh Tran. 2019. "Nationalizing a Global Phenomenon: A Study of How the Press in 45 Countries and Territories Portrays Climate Change." *Global Environmental Change : human and Policy Dimensions* 58 (April): 101942. doi:10.1016/j.gloenvcha.2019.101942.

Walter, Dror, and Yotam Ophir. 2019. "News Frame Analysis: An Inductive Mixed-Method Computational Approach." *Communication Methods and Measures* 13 (4): 248–266. doi:10.1080/19312458.2019.1639145.

Wilkerson, John, and Andreu Casas. 2017. "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges." *Annual Review of Political Science* 20 (1): 529–544. doi:10.1146/annurev-polisci-052615-025542.

Willis, Rebecca. 2017. "Taming the Climate? Corpus Analysis of Politicians' Speech on Climate Change." *Environmental Politics* 26 (2): 212–231. doi:10.1080/09644016.2016.1274504.