Constructions: a new unit of analysis for corpus-based discourse analysis

Samia Touileb
Information Science and Media Studies
University of Bergen
N-5020 Bergen
Norway
samia.touileb@gmail.com

Andrew Salway
Uni Research Computing
Thormøhlensgt. 55
N-5008 Bergen
Norway
andrew.salway@uni.no

Abstract

We propose and assess the novel idea of using automatically induced constructions as a unit of analysis for corpus-based discourse analysis. Automated techniques are needed in order to elucidate important characteristics of corpora for social science research into topics, framing and argument structures. Compared with current techniques (keywords, n-grams, and collocations), constructions capture more linguistic patterning, including some grammatical phenomena. Recent advances in natural language processing mean that it is now feasible to automatically induce some constructions from large unannotated corpora. In order to assess how well constructions characterise the content of a corpus and how well they elucidate interesting aspects of different discourses, we analysed a corpus of climate change blogs. The utility of constructions for corpus-based discourse analysis was compared qualitatively with keywords, n-grams and collocations. We found that the unusually frequent constructions gave interesting and different insights into the content of the discourses and enabled better comparison of sub-corpora.

1 Introduction

In recent years, with the increasing availability of online text data and computing power, there has been a rapid increase in interest in corpus-based discourse analysis, particularly among social science researchers. Within social science, discourse analysis is concerned with how societally important issues and opinions are expressed through language, e.g. in news and social media. The scale of the data sets means that automated techniques are essential, at least to give researchers an overview of the content in a corpus and to elucidate interesting aspects for further investigation.

The aim of this paper is to assess the novel idea of using automatically induced constructions for corpus-based discourse analysis. Section 2 provides some background about corpus-based discourse analysis and discusses some limitations of the automated techniques that are commonly used. It also describes what constructions are and how some constructions can be induced automatically by taking advantage of recent developments in natural language processing. Then in Section 3 we report our investigation into the use of constructions for corpus-based discourse analysis. This compared the utility of unusually frequent constructions with current techniques, based on how they gave insights into the content of a large corpus of climate change blogs, and how they elucidated interesting phenomena for further investigation. Section 4 summarises our conclusions and contributions, and outlines future work.

2 Background

In this section we review the use of automated text analysis techniques for corpus-based discourse analysis, and explain why we propose constructions as a new unit of analysis (section 2.1). Then we explain how the state-of-the-art in grammar

28th Pacific Asia Conference on Language, Information and Computation pages 634–643

induction means that it is now possible to automatically induce some constructions from unannotated corpora (section 2.2).

2.1 Corpus-based discourse analysis

In the social sciences, the term discourse is used to refer to how ideas and opinions are formed, influenced and expressed through language (Baker, 2006). Researchers study discourses in order to explain the effect of language use on social, political, legal and environmental issues, among many others. An often-cited and simple example is how the difference between referring to an individual as a "freedom fighter" or a "terrorist" effects a reader's perception and opinions.

Corpus-based approaches take advantage of automated techniques in order to analyse large-scale discourses such as those in corpora of news and social media (e.g. Fløttum et al, 2014; Kim, 2014; Jaworska and Krishnamurthy, 2012; Grundman and Krishnamurthy, 2010). The techniques can reveal interesting phenomena within the corpus that would not be apparent to a researcher who read the material (Baker, 2006); often there is too much material for a researcher to read anyway. That said, automated analyses alone are not normally sufficient: they must be complemented with manual inspections of the texts and consideration of their contexts.

For many social science researchers, an important part of discourse analysis is the characterisation of how issues are framed. To frame an issue is to "select some aspects of a perceived reality and make them more salient in a communicating text" (Entman, 1993). Framing is also defined as "a central organizing idea or story line that provides meaning to an unfolding strip of events" (Gamson and Modigliani, 1989).

Framing analysis necessarily involves text analysis in order to identify salient formulations (frames) and to uncover how issues are represented differently by participants in discourses. In a recent paper, Touri and Koteyko (2014) provide an extensive review of methods for framing analysis and describe ways in which corpus linguistic techniques can be applied, with a focus on keywords and concordances. A keyword list helps to identify words that indicate what perspective is being taken on an issue; cf. the "freedom fighter/terrorist" example. Then, concordances which show instances of words and their co-texts can be read in order to understand more about the ways in which words are being used as parts of frames.

Another recent paper shows how collocation data can be used to analyse how issues are represented in the media (McEnery et al., 2013). Statistically significant collocations around words that refer to an issue of interest are interpreted, for example, as giving a positive or negative tone.

There is also the potential for automated techniques to contribute to investigations in other areas of social science research by identifying some of the linguistic patterns that are used to build discourses. For example, the ability to characterise and compare dominant topics, and the ways in which they are expressed, is relevant for investigating: agenda setting – what issues get more attention in the media, e.g. (Grundman and Krishnamurthy, 2010); polarisation – how different social groups form increasingly divergent opinions, e.g. (Elgesem et al., 2014), (Adamic and Glance, 2005); and argument structures – the ways in which writers try to persuade others, e.g. (Koteyko et al., 2013).

In general, keywords and n-grams can be seen as highlighting salient ideas and opinions in discourses. Collocations characterise language use around keywords and can be seen as giving insights into the meanings typically associated with issues. However, as noted previously, these techniques can only be a starting point for a researcher. The lack of information about the co-text around keywords and n-grams restricts the extent to which they can be interpreted without the close reading of concordances. Increasingly, corpora of interest to social scientists are too large for close reading of all the relevant concordances, so we see a need for techniques to condense information about co-texts.

Collocation data already provides some information about a keyword's co-text, i.e. it shows the words that have a statistically significant association with the keyword. However, collocation data is typically presented as a large grid of statistics for one keyword. It seems to us that it would be desirable to have a simpler picture that is more intuitive to interpret.

Furthermore, by prioritising lexical elements, the use of keywords, n-grams and collocations may fail to elucidate relevant grammatical phenomena. As noted by Baker (2006), unusually frequent grammatical phenomena (as well as words and phrases), can also reveal the non-obvious meanings

of a discourse. It seems to us that they are particularly important for framing and argumentation analysis.

All these observations lead us to propose constructions as a new unit of analysis for corpus-based discourse analysis, to complement existing techniques. A construction is defined as a form-meaning pair (Goldberg, 2009). The form of a construction can be any combination of morphemes, words, phrases, idioms, local grammatical templates and word classes, as well as general linguistic structures. Thus, we see constructions as a convenient way to conceptualise language for the purposes of corpus-based discourse analysis. Firstly, they encompass a wide variety of linguistic forms. Secondly, these forms are thought of as mapping directly to meaning which is the ultimate object of study in discourse analysis. In particular, constructions that capture local grammatical templates and word classes may be particularly useful.

Our idea is that a researcher can start an investigation by looking at a set of salient constructions, perhaps alongside keywords, n-grams and collocations, in order to get deeper insights into the distinctive characteristics of a particular discourse. In the following sub-section we discuss how it is possible to induce some salient constructions automatically from an unannotated corpus.

2.2 The automatic induction of constructions

Developments in natural language processing have led to the automatic induction of grammatical structures from unannotated corpora, e.g. the ADIOS algorithm (Solan et al., 2005); see D'Ulizia et al. (2011) for a review of the field of grammatical inference.

ADIOS (Automatic DIstillation of Structure) is an unsupervised algorithm that discovers hierarchical structures in sequential data, e.g. words in sentences. It identifies the most significant patterns (horizontal sequences) and equivalence classes (vertical groups) within the context of patterns, using statistical information. Each sentence is loaded onto a directed pseudograph with one vertex for each vocabulary item: this means that partially aligned sentences share sub-paths across the graph. In each iteration, the most significant pattern is identified with a statistical criterion that favours frequent sequences that occur in a variety of contexts. Then, the algorithm looks for possible equivalence classes within the context of the pattern, i.e. it identifies positions in the pattern that could be filled by different items and forms an equivalence class with those items. At the end of the iteration, the new pattern and equivalence class become vocabulary items in the graph, so that they can become part of further patterns and equivalence classes, and hence hierarchical structures are formed.

From our point of view, ADIOS has three particularly good features. Firstly, it is unsupervised which means that it should be portable across different languages and domains. Secondly, since equivalence classes only exist in the specified contexts of patterns, the structures induced by ADIOS will generate less overgeneralization than methods assigning global categories to each unit of a sentence, i.e. it gives a better description of local grammatical features. Thirdly, induced patterns may encapsulate units occurring in positions far apart from each other.

The ADIOS algorithm, like some others, builds on the insights of Zellig Harris who argued that grammatical structures can be induced through a distributional analysis of the surface forms of languages (Harris, 1954). He also showed how linguistic structures that are identified in this way map to important information structures, especially in domain-specific corpora (Harris, 1988).

This second point motivated work to modify and apply the ADIOS algorithm for text mining purposes, i.e. to extract salient information structures from an unannotated corpus (Salway and Touileb, 2014). The learning regime of ADIOS was modified in order to focus the algorithm on text snippets around key terms of interest, rather than processing all sentences. This change was influenced by the theory of local grammars (Gross, 1997), i.e. the idea that language is best described with word classes that are specific to local contexts. Another modification targeted the most frequent and meaningful structures. To do this, after each iteration, instances of the most frequent patterns were replaced with common identifiers in the input file so that patterning around them was more explicit in subsequent iterations.

Following this method, 671 patterns were induced from a corpus of climate change blogs by Salway and Touileb (2014); see section 3.1 for a description of this corpus. Table 1 shows some examples of the patterns generated by the automatic process. The patterns and the equivalence clas-

ses that they contain are bracketed. The elements of patterns are separated by white space and the elements of equivalence classes are separated by '|'.

Pattern 1 in the table captures a simple word sequence which is a domain term – "fossil fuels". Pattern 2, with the equivalence class "(carbon|(greenhouse gas)|co2)", captures three near-equivalent domain terms – "carbon emissions", etc. Pattern 3 does something similar to capture two interchangeable phrases that are common in the corpus; note, in this pattern there is overgeneralization due to the equivalence class "(of|for)". Pattern 4 shows some grammatical structure being captured with three verbs – "(combat|minimize|tackle)" – that appeared in the same context in the corpus. Patterns 5 and 6 capture both grammatical structure and some near-synonyms.

1. (fossil fuels)
2. ((carbon|(greenhouse gas)|co2) emissions)
3. ((consequences|impacts) ((of|for) climate change))
4. ((to (combat|minimize|tackle)) climate change)
5. (((due to)|(caused by)) ((climate change)|(global warming)))
6. (((((global|some|sophisticated|complex|the) climate models)|climate models) (project|suggest|predict)) that)
7. ((of global warming) (was|are|is))
8. (in (order|(the (atmosphere|recessions))))

Table 1. Examples of the patterns induced from a corpus of climate change blogs (Salway and Touileb, 2014).

Given Goldberg's definition of a construction, cf. section 2.1, it seems reasonable to refer to patterns 1-5 as constructions. Of course, that is not to say that the induction process captures all kinds of constructions. Rather, it seems to capture mainly terms, phrases and local grammatical templates. We previously noted the need for techniques to condense information about keywords' co-texts, in order to reduce the need for reading large quantities of concordance lines. It may be argued that patterns 3-5 are doing a useful job in condensing some of the co-texts around "climate change".

It should be noted that some patterns are incomplete constructions, e.g. "7. ((of global warming) (was|are|is))", and others are not constructions at all because they mix grammatical structures and contain equivalence classes that are semantically incoherent, e.g. "8. (in (order|(the (atmosphere|recessions))))".

Since we have no automatic way to separate patterns that are constructions from those that are not constructions, we can only use the complete set of patterns for corpus-based discourse analysis, cf. section 3.2. As will be seen in section 3.3, the presence of patterns that are not constructions does not have an adverse effect on results. For convenience, from this point forward, we refer to the set of patterns as a set of constructions, whilst noting that it contains some non-constructions.

3 Assessing the use of constructions for corpus-based discourse analysis

The investigation focussed on two main questions. (1) Do unusually frequent constructions reflect the distinctive content of a (sub-) corpus? (2) If so, do they suggest interesting lines of further investigation for discourse analysis?

In order to answer these questions, we analysed constructions in a corpus of climate change blogs. Specifically, we identified unusually frequent constructions in three major blogs (which can be considered as sub-corpora), and qualitatively evaluated the utility of these constructions for corpus-based discourse analysis. We then compared their utility with keywords, n-grams and collocations.

Section 3.1 describes the climate change corpus and the three blogs analysed. Section 3.2 describes how unusually frequent constructions were identified. Section 3.3 discusses how these constructions give insights into the content of each blog and how they suggest further lines of investigation for corpus-based discourse analysis. Section 3.4 compares the insights gained from the constructions with what can be learnt from keywords, n-grams and collocations for the same blogs. Section 3.5 discusses the findings with respect to the two questions stated above.

3.1 Corpus

The NTAP corpus comprises about 3000 English language blogs (1.4 million blog posts) related to climate change issues (Salway et. al, 2013). This corpus is interesting for discourse analysis because climate change is a complex and contested issue with diverse sub-topics, perspectives and opinions. It may be hypothesised that the discourses around

climate change are polarized (sceptics and acceptors), framed in different ways (e.g. science, politics, national and local issues), and contain a variety of argumentation structures used to support different positions.

As an example of social media, blogs represent both an opportunity and challenge for corpus-based discourse analysis. They may reflect a greater variety of perspectives and opinions than traditional media. However, the large volume of material and the greater variety of language use mean that new unsupervised automated techniques are required.

For assessing the utility of unusually frequent constructions, we focussed our analysis on three major blogs that we already knew something about (Elgesem et al., 2014). The blog *wattsupwiththat.com* (4996 posts; 3.5m words) is one of the most central blogs in the sceptical blog community and is concerned with climate science issues. The blog *itsgettinghotinhere.org* (1343 posts; 0.8m words) is a central blog in the accepters community and discusses both climate science and climate politics. The third blog, *chimalaya.org* (3782 posts; 3.1m words) has many links to the other two blogs, and is concerned with climate politics issues for the Himalaya region.

3.2 Unusually frequent constructions

We took the set of constructions extracted by Salway and Touileb (2014), as described in section 2.2; recall, this set includes some patterns that are not constructions but we refer to it as a set of constructions for convenience. It was decided that constructions with frequency less than 50 in the whole corpus were unlikely to be unusually frequent in any single blog and so they were removed. Then we counted the frequency for each remaining construction (381 constructions) in each of the three blogs. This was straightforward because each construction is described as a regular expression.

In order to identify the unusually frequent constructions in each blog relative to the other two blogs, we used the RRF statistic – ratio of relative frequencies (Edmundson and Wyllys, 1961). This is a simple measure that reflects how much more (or less) something appears in corpus A compared to corpus B, whilst factoring in the sizes of the corpora. The RRF for a unit is computed as:

$$RRF_U = {RF_{UA}}/{RF_{UB}}$$

RF_{UA}: Relative frequency of unit U in corpus A.
RF_{UB}: Relative frequency of unit U in corpus B.

Where:

$$RF_U = {F_U}/{N}$$

F_U: Frequency of unit U in the corpus.
N: Total number of words (tokens) in the corpus.

Note, there can be an issue with division by zero in the RRF equation when F_U is zero in corpus B. However this situation did not arise in the current analysis.

For each of the three blogs we ranked the 381 constructions according to their RRF values, where corpus B was the union of the other two blogs. The RRF statistic can give misleading results for low frequency values: it is "easier" for a low-frequency item to get a high RRF value. With this in mind, a frequency threshold was applied to the ranked lists of constructions. After testing various thresholds, it was decided to use a frequency threshold equal to 0.001% of the size of each blog. Thus constructions only appear in the ranked RRF lists if they have frequencies greater than: *chimalaya* (30), *itsgettinghotinhere* (8), *wattsupwiththat* (34). These thresholds mean that we can be more confident that the ranked constructions for a blog are reflective of that blog's content in general, rather than just a few blog posts within it.

3.3 Results

Table 2 presents the top 10 constructions ranked by RRF values for the three blogs *chimalaya*, *itsgettinghotinhere* and *wattsupwiththat*. These are the most unusually frequent constructions that we assume will reveal some of each blog's distinctive characteristics. Each construction is presented with an ID (for ease of reference), and using brackets and '|'s as described in section 2.2. For each construction the table gives its total frequency, and then a breakdown of the frequencies of its various forms. For example, C2 (C for *chimalaya*) occurs 1172 times in total – 1061 times as "developing countries" and 111 times as "poor countries".

We envisage a social science researcher using ranked lists of constructions as a starting point to investigate the discourses in one or more (sub-) corpora. Thus, the constructions should provide a convenient overview of the content and draw atten-

tion to potentially interesting phenomena, like topics, framing and argument structures. In the following sub-sections we discuss how the constructions in Table 2 could be used for these purposes.

3.3.1 Constructions elucidating topics?

Many of the constructions in Table 2 do indeed reflect what we already know about the content of the blogs: *chimalaya* – climate politics, Himalaya region; *itsgettinghotinhere* – climate science, climate politics; *wattsupwiththat* – sceptical views of climate science. Furthermore, many of the constructions give a finer-grained view on how the distinctive topics are expressed in each blog.

For example, constructions C1, C3, C5 and C9 all indicate that *chimalaya* focusses on the impacts/effects of climate change, rather than its causes. Constructions C3 and C9 include both "causes" and "effects" but from the frequencies of the different forms it is apparent that this blog is much more concerned with the effects. The blog's interests in addressing climate change are highlighted by constructions C7 and C8, with frequent mentions of meetings in C4. Its focus on the kinds of countries that comprise the Himalaya region is indicated by C2.

Itsgettinghotinhere's constructions I5, I6 and I9 all highlight its concern with taking action to address climate change issues, although perhaps cotexts for I5 and I9 should be checked to confirm this. Constructions I1, I4, I7 and I8 are terms that suggest a focus on discussing the link between climate change and energy production. Various ways to express the idea of "cap and trade schemes" as part of a solution to climate change are captured by I2, and partially by the incomplete construction I3.

Constructions W3 and W10 indicate that *wattsupwiththat* discusses the role of humans in causing global warming, although none of the constructions indicate this blog's sceptical viewpoint, except perhaps the form "no global warming" in W10. The partial constructions W4 and W8 suggest an interest in climate models, but further investigation would be needed to see what is being said about them. Compared with the other two blogs, we get a less clear picture of this blog's distinctive content.

3.3.2 Constructions related to frames?

As discussed in section 2.1, framing analysis has benefited from automated techniques such as keywords and collocations. However, we noted the potential for constructions to elucidate richer linguistic patterning that could be related to how different perspectives are represented in corpora. Here we give some examples of how constructions highlight framing phenomena that would not be so apparent using current techniques.

It could be argued that the construction "C2 ((developing|poor) countries)" suggests that in *chimalaya* the climate issue is framed from the perspective of developing countries and their particular concerns. We note though that, in this case, there is a fuzzy boundary between this notion of framing and the notion of topic. A clearer framing interpretation is the strong preference for the form "developing countries" (f=1061) compared with "poor countries" (f=111) which indicates a choice to frame these countries in a positive way.

Another interesting construction that is unusualy frequent in this blog is "C8 (to (combat|minimize|tackle)) climate change)". The construction itself suggests two different framings on how the climate issue can be addressed. Firstly, there is a rather dispassionate and diplomatic approach – indicated by the form "to minimize climate change". Secondly, there is a more passionate and confrontational position which is expressed with stronger words – "to combat|tackle climate change". The frequencies of these forms within *chimalaya* make it clear that this blog is firmly taking the second position (f=1 vs f=129); this is further supported by C7. Perhaps collocation data would show "combat" and "tackle" as being associated with "climate change" in this blog: however, the grammatical structure captured by C8 also elucidates the contrast with "minimize".

The construction W3, which is unusually frequent in *wattsupwiththat*, highlights a difference in framing between saying "man made global warming" and "anthropogenic global warming". Whilst these terms have the same meaning, the latter has a more scientific connotation. The preference for the form "anthropogenic global warming" in this blog strikes us as interesting, because in another analysis we have seen a general preference for "man made global warming" in sceptical blogs. This prompted us to look at the concordances for

chimalaya.org

C1. (impact ((of|for) climate change)): 284 - *impact of climate change (*284)

C2. ((developing|poor) countries): 1172 - *developing countries(1061), poor countries (111)*

C3. (the (causes|effects) | (consequences|impacts) ((of|for) climate change))): 460 - *the impacts of climate change (224), the effects of climate change (203) , the consequences of climate change (29), the causes of climate change (4)*

C4. (climate change (talks|meeting|summit| conference)): 131 - *climate change conference (55), climate change talks (47), climate change summit (21)*

C5. ((consequences|impacts) ((of|for) climate change)): 478 - *impacts of climate change (416), consequences of climate change (62)*

C6. (\d+ per cent): *695* - *\d+ per cent (695)*

C7. (tackling climate change): 47 - *tackling climate change (47)*

C8. ((to (combat|minimize|tackle)) climate change): 130 - *to tackle climate change (72), to combat climate change (57), to minimize climate change (1)*

C9. ((causes|effects) ((of|for) climate change)): 357 - *effects of climate change (345), causes of climate change (12)*

C10. (to climate change): 1289 - *to climate change (1289)*

itsgettinghotinhere.org

I1. (global warming pollution): 23 - *global warming pollution (23)*

I2. (a (cap and) ((trade|trading|cap and trade) (scheme|system|program|approach))): 13 - *a cap and trade system (8), a cap and trade program (4), a cap and trade scheme (1)*

I3. (cap and): 92 - *cap and (92)*

I4. (clean air): 33 - *clean air (33)*

I5. (to (stem|stop)): 266 - *to stop (263), to stem (3)*

I6. (action (on climate change)): 36 - *action on climate change (36)*

I7. (power plants): 133 - *power plants (133)*

I8. (fossil fuels): 213 - *fossil fuels (213)*

I9. (to regulate): 29 - *to regulate (29)*

I10. (a (pilot|national|possible|nationwide|broad based)): 108 - *a national (97), a nationwide (6), a possible (3), a pilot (2)*

wattsupwiththat.com

W1. (the carbon tax): 37 - *the carbon tax (37)*

W2. ((global warming|((and|to) global warming)) (has|can|will)): 71 - *global warming has (32), global warming will (27), global warming can (8), and global warming has (2), to global warming will (2)*

W3. ((man made|anthropogenic) global warming): 69 - *anthropogenic global warming (61), man made global warming (8)*

W4. ((analysing|in|on|by) climate models): 45 - *in climate models (24), by climate models (15)*, *on climate models (6)*

W5. (global warming (is|was)): 226 - *global warming is (200), global warming was (26)*

W6. ((to|between|by|about) \d+): 4382 - *to \d+ (1985), about \d+ (1363), by \d+ (659), between \d+ (375)*

W7. ((would|will) be): 3239 - *will be (1873), would be (1366)*

W8. ((global|some|sophisticated|complex|the) climate models): 126 - *the climate models (92), global climate models (25), some climate models (4), complex climate models (4), sophisticated climate models (1)*

W9. ((who|he) (was|are|is)): 915 - *he was (248), who are (203), he is (189), who is (169), who was (106)*

W10. ((a|no) (((man made|anthropogenic) global warming)|global warming)): 40 - *a global warming (22), no global warming (18)*

Table 2. Top 10 constructions ranked by RRF for three blogs. Each construction is given with ID, its total frequency, and the frequencies of its different forms.

"anthropogenic global warming" within *wattsupwiththat*. We saw that it was typically used to frame the issue in scientific terms, but then to comment on the views of climate scientists in negative and sarcastic ways.

3.3.3 Constructions related to argument structures?

The construction "W7 ((would|will) be)" struck us as interesting because it contains only grammatical words. Since these words are usually very frequent and part of general language, it is particularly interesting when they have a high RRF. By looking at the frequencies of the two forms of W7 in the three blogs, we see that its high RRF is mainly due to a relatively high use of the form "would be". In the other two blogs the frequency of "would be" is less than 45% of the frequency of "will be", but in *wattsupwiththat* it is 73%, Table 3.

Blog	"will be"	"would be"
wattsupwiththat	1873	1366
chimalaya	2122	891
itsgettinghotinhere	564	250

Table 3. Frequencies of the forms of W7.

From a preliminary analysis of the concordances of "would be" in *wattsupwiththat*, we got the impression that it is being used as part of argumentation structures in a scientific style of language; for example, statements of hypotheses like "if X then Y would happen". This could perhaps be a starting point for investigating the degree to which climate issues are discussed in a scientific style across the blogosphere.

Another example of a construction that relates to argument structures was found just outside of the top 10: this was "((you|we) (can|should))" which was 15th in the ranking for *itsgettinghotinhere*. The frequencies of its four forms were: "we can" (f=302), "you can" (f=196), "we should" (f=84), "you should" (f=8). The preference for "we" versus "you" suggests that the writers are trying to be inclusive of their readers, and are urging for collective action against climate change. This perhaps contrasts with the third person style of scientific writing in other blogs.

The even stronger preference for "can" versus "should" suggests that the writers are trying to maintain an encouraging and positive tone, and to avoid alienating people by not telling them directly what to do. Of course, all these observations would have to be supported by more analyses, but it seems that the constructions did highlight interesting aspects of the discourses.

3.4 Comparison with current techniques

In order to make a qualitative comparison between the use of constructions and current techniques, we generated keyword, n-gram and collocation data from the same three blogs. Of course, there are multiple ways to implement these techniques so a comprehensive comparison is not possible here. We have tried to follow typical implementations of the techniques and believe that our general observations would hold regardless of implementation details. We recognise the need for more extensive and quantitative evaluation in future work, but this was beyond the scope of the current paper.

3.4.1 Keywords and key n-grams

We generated a list of 20 keywords and 20 key n-grams for each blog, using a frequency threshold and the RRF statistic to rank them, cf. section 3.2. Some of *chimalaya's* keywords and n-grams reflect the fact that it is broadly about climate and the Himalaya region, e.g. "Kashmir", "Nepalese", "Bhutanese", "Punjab", "GEF" (Global Environment Facility), "in the Himalayan region", "mountain ecosystem", "climate related issues". There are also indications of its interest in development, e.g. "ADB" (Asian Development Bank), "knowledge sharing", "capacity building".

Similarly, some keywords and key n-grams point broadly to the topics of the other two blogs: *itsgettinghotinhere* – "BP" (British Petroleum), "RBC" (Royal Bank of Canada), "clean energy economy", "action network"; *wattsupwiththat* – "OHC" (Ocean Heat Content), "ASOS" (Automated Surface Observing Stations), "MMTS" (Maximum/Minimum Temperature System), "linear trend", "data sets", "climate audit".

It might be possible to use some of the keywords and n-grams as the starting point for framing analysis, cf. the method described by Touri and Koteyko (2014). However this would entail extensive reading of concordance lines. On a separate point, as far as we can see, none of the keywords and n-grams suggest distinctive argument structures.

3.4.2 Collocations

We generated a list of the top 10 collocates of the word "climate" in each blog, using a span of +/- 5 words, and ranking on mutual information (Baker, 2006); again the 0.001% frequency threshold was applied.

In all three blogs there was an unsurprising association between "climate" and "change". More specifically, in c*himalaya* the words most strongly associated with "climate" included "intergovernmental" and "panel" which point to the term "Intergovernmental Panel on Climate Change". Other strongly associated words point to the blog's interest in addressing climate change, e.g. "combat", "combating", "adapting", "mitigating". Likewise, collocates of "climate" in the other two blogs also reflected something about their foci: *itsgettinghotinhere* – "causes", "effects", "impact", "addressing"; *wattsupwiththat* – "denier", "impacts", "panel", "framework", "intergovernmental".

3.5 Discussion

The results from this investigation suggest that a list of unusually frequent constructions reflects some of the distinctive content of a (sub-) corpus. Further, and in answer to our second question, there were examples of constructions that revealed linguistic patterning that would be of interest for further analysis into topics, framing and argumentation structures.

With regards to topic analysis, the constructions are useful because, unlike keywords, they capture terms and phrases which could enable finer-grained topic classification and text retrieval. Terms and phrases will be present in n-gram lists but these lists are typically very long and noisy. A further apparent advantage of constructions is that they group together alternative ways to refer to the same concept.

For the analysis of framing and argumentation structures, the fact that some constructions explicate local grammatical structures gives an advantage over current techniques. For example, the construction "(to (combat|minimize|tackle)) climate change)" highlights a potential framing choice more explicitly than the equivalent keyword or collocation data. The words "combat", "minimize" and "tackle" could appear as keywords and collocates, but the researcher would have to then analyse large numbers of concordance lines to establish that they were part of frames.

It was also seen that some constructions comprising only grammatical words highlighted linguistic patterning that was relevant for the analysis of argument structures, i.e. "((would|will) be)" and "((you|we) (can|should))". The grammatical structures in these constructions would certainly not be apparent with current techniques, and indeed it is unlikely that the individual words would even be noticed in lists of keywords and collocates because they are so frequent in general language.

4 Concluding remarks

This paper has proposed and assessed the novel idea of using constructions as a unit of analysis for corpus-based discourse analysis. We envisage researchers consulting lists of unusually frequent constructions as a first step in data-driven investigations, i.e. in order to get an overview of the content of large corpora, and to identify interesting phenomena for more detailed analysis. The use of constructions is appealing because, unlike current techniques, they capture both lexical and grammatical patterning.

Building on recent work in natural language processing it was possible to automatically identify unusually frequent constructions within a large corpus of climate change blogs. We showed how lists of unusually frequent constructions highlighted a variety of linguistic phenomena relating to topic, framing and argumentation structures. These phenomena would all be interesting for corpus-based discourse analysis and would not be so apparent to researchers using keywords, n-grams, collocations and concordances.

Whilst we only looked at constructions within one corpus, there is good reason to believe that the approach would be broadly applicable because the induction process is unsupervised. That said, because the induction process exploits partially overlapping word sequences around key terms, we expect that it will be most effective on large corpora with relatively constrained language use. In other words, it will work best with corpora that consist of a single domain and a single text genre.

In order for this approach to be integrated into social science research methods, it will be important to understand more about how the induction process works. Although we can observe the

interesting constructions that it gives, as yet we know little about what it misses and why. See Salway and Touileb (2014) for more about related ongoing work. This must include a more rigorous, and ideally automated, separation of induced patterns into constructions and non-constructions.

Acknowledgments

We thank Zach Solan for providing an implementation of the ADIOS algorithm, Knut Hofland and Lubos Steskal for creating the NTAP blog corpus, and Dag Elgesem for his helpful comments. This research was supported by a grant from The Research Council of Norway's VERDIKT program.

References

Lada Adamic and Natalie Glance. 2005. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. *Procs. of the 3rd International Workshop on Link discovery - LinkKDD*, 36–43.

Paul Baker. 2006. *Using corpora in discourse analysis*. 2006 London: Continuum.

Arianna D'Ulizia, Fernando Ferri and Patrizia Grifoni. 2011. A survey of grammatical inference methods for natural language learning. *Artificial Intelligence Review,* 36(1):1-27.

Harold P. Edmundson and Ronald Eugene Wyllys. 1961. Automatic Abstracting and Indexing - Survey and Recommendations. *Communications of the Association for Computer Machinery*, 4(5).

Dag Elgesem, Lubos Steskal and Nicholas Diakopoulos. 2014. The structure and content of the discourse on climate change in the blogosphere: the big picture. To appear in: *Environmental Communication. Special issue on climate change communication on the Internet.*

Robert Entman. 1993. Towards clarification of a fractured paradigm. *Journal of Communication*, 43(4):51-58.

Kjersti Fløttum, Øyvind Gjerstad, Anje Müller Gjesdal, Nelya Koteyko and Andrew Salway. 2014. Representations of the FUTURE in English language blogs on climate change. To appear in: *Global Environmental Change.*

William A. Gamson and Andre Modigliani. 1989. Media discourse and public opinion on nuclear power: a constructionist approach. *American Journal of Sociology*, 95(1):1-37.

Adele Goldberg. 2009. The nature of generalization in language. *Cognitive Linguistics*, 20(1):93–127.

Maurice Gross. 1997. The Construction of Local Grammars. In: E. Roche and Y. Schabes (eds.), Finite-State *Language Processing. The MIT Press, Cambridge MA*, 329-354.

Reiner Grundmann and Ramesh Krishnamurthy. 2010. The Discourse of Climate Change: A Corpus-based Approach. *Critical Approaches to Discourse Analysis Across Disciplines*, 4(2):125-146.

Zellig Harris. 1954. Distributional structure. *Word*, 10:(2/3).146-162.

Zellig Harris. 1988. *Language and Information*. Columbia University Press, New York.

Sylvia Jaworska and Ramesh Krishnamurthy. 2012. On the F word: A corpus-based analysis of the media representation of feminism in British and German press discourse, 1990–2009. *Discourse & Society*, 23(4):401-431.

Kyung Hye Kim. 2014. Examining US news media discourses about North Korea: A corpus-based critical discourse analysis. *Discourse & Society*, 25(2):221-244.

Nelya Koteyko, 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.

Tony McEnery, Amanda Potts and Richard Xiao. 2013. Is there a reputational benefit to hosting the Olympics and Paralympics? *Procs. Corpus Linguistics 2013,* Lancaster University.

Andrew Salway, Knut Hofland and Samia Touileb. 2013. Applying Corpus Techniques to Climate Change Blogs. *Procs. Corpus Linguistics 2013,* Lancaster University.

Andrew Salway and Samia Touileb. 2014. Applying grammar induction to text mining. *Procs. 52nd ACL Conference* (short papers)*,* 712-717.

Zach Solan, David Horn, Eytan Ruppin and Shimon Edelman. 2005. Unsupervised learning of natural languages. *Procs. of the National Academy of Sciences,* 102(33):11629–11634.

Maria Touri and Nelya Koteyko. 2014. Using corpus linguistic software in the extraction of news frames: towards a dynamic process of frame analysis in journalistic texts. *International Journal of Social Research Methodology*, Published online 3 July2014. DOI:10.1080/13645579.2014.929878.