Pre-print version.

Salway, A. 2017. Data-driven approaches to climate change discourse, illustrated through case studies of blogs and international climate negotiation. In K. Fløttum ed., *The Role of Language in the Climate Change Debate*, 151-170. New York and London: Routledge.

Data-driven approaches to climate change discourse, illustrated through case studies of blogs and international climate negotiation

Introduction

Observations about the increasing volume and diversity of textual material available for humanities and social science research are at the tipping point of cliché but nevertheless point to exciting and challenging opportunities. The climate change debate manifests in a great diversity of texts – in multiple genres and many languages across the globe – produced by a wide variety of actors concerned with different aspects of climate change and wishing to promote different points of view. In addition to scientific reports, policy documents, and newspaper stories which have been quite well studied, recent review papers highlight the need to pay more attention to social media data such as blogs and microblogs like Twitter (Schäfer 2012; Schäfer and Schlichting 2014; Auer, Zhang, and Lee 2014). On top of already large amounts of news media, the widespread uptake of social media means that the volume of textual material relating to climate change is vast and increasing rapidly. Recent collections of English language texts alone suggest 30,000 newspaper articles, 2.8 million tweets, and 150,000 blog posts per month broadly related to climate change (Olteanu et al. 2015; Salway et al. 2016).

Compare these numbers to the sample sizes that are feasible when manual methods are used to investigate climate change discourse, for example: manually coding frames in 974 newspaper stories (Boykoff 2008); identifying eight distinct "discourses" from about 150 newspaper stories (Doulton and Brown 2009); analyzing key themes in 3 blogs based on reading of 20 blog posts from each (Sharman 2014); and, analyzing frames in 300 tweets (O'Neill et al. 2015). It is apparent that automatic text analysis techniques must be incorporated into research methods in order to deal with the volume of climate change texts. Of course, it is crucial to consider the

trade-off between the depth, richness, and sensitivity of analysis that is possible for a human analyst and the breadth of coverage that automated techniques enable. At the very least however, automation has a role in identifying texts that are related to the topic of climate change: e.g. the use of queries to identify relevant texts for investigating the coverage of climate change in news media and social media, both over time and geographically (Boykoff and Roberts 2007; Schmidt, Ivanova, and Schäfer 2013; Grundmann and Scott 2014; Kirilenko, Molodtsova, and Stepchenkova 2015; Olteanu et al. 2015). Beyond that, the focus in this chapter is on the use of automated techniques to analyze the content of textual data sets, known as corpora, related to climate change, i.e. to provide manageable overviews of the content, to highlight interesting linguistic patterning and to automatically annotate, or code, the material. I argue that the scale, complexity, variety, and dynamic nature of climate change discourses mean that data-driven approaches in particular have an important role in analyzing large text corpora. In brief, data-driven approaches may help to minimize biases due to prior assumptions about the form and content of the textual material, and to generate new insights and stimulate hypotheses.

The next section discusses what data-driven means specifically in relation to automatic text analysis and sketches a general approach for data-driven research. This is followed by a brief review of the text analysis techniques that have been used to investigate climate change discourse, especially in news media and social media, i.e. text classification, topic modelling, and various corpus linguistic techniques that have been used for corpus-based discourse analysis. The review concludes that corpus-based discourse analysis is important in order to understand texts as more than "bags of words", i.e. to not only treat language at the lexical level, and that new techniques are required to enable researchers to identify linguistic patterning in large corpora. These points are then explored and developed through three case studies which reflect

on the data-driven text analysis techniques that were incorporated into corpus-based methods for investigating climate change discourse. The latter two case studies made use of a recently developed technique for elucidating linguistic patterning – local grammar induction – which is introduced in its own section.

What does "data-driven" mean for automatic text analysis?

Automatic text analysis, in essence, involves counting instances of linguistic forms that have meaning with respect to a conceptual framework that, in turn, articulates aspects of phenomena that are the object of an investigation, e.g. topics, media representations, frames, rhetorical strategies, and ideologies. A simple case would be the counting of words associated with different topics in order to analyze the topics' prevalence across a corpus of texts. Thus, when considering the relative merits of different approaches to text analysis we should pay attention to how decisions are made in relation to: (i) selecting non-textual phenomena for investigation; (ii) determining an appropriate conceptual framework; (iii) establishing a mapping between concepts and linguistic forms that can be counted, i.e. a coding scheme; and, (iv) identifying interesting and significant statistical results in the frequency counts of linguistic forms.

A premise of the current chapter is that the increasing volume and diversity of textual material related to climate change mean that data-driven approaches will become increasingly relevant, and perhaps essential, as part of each of (i)-(iv) above, and especially (iii). The diversity of material from different domains, comprising various text genres, and in multiple languages, means that the situation in which researchers are faced with textual material that they are not familiar with will become commonplace. Thus it will be harder for researchers to rely on previously developed theory, conceptual frameworks, and coding schemes to guide and implement their investigations. As for volume, when text corpora already comprise 100s and

1000s of millions of words, even if a researcher is only interested in one particular word or term then they may be faced with 100,000s of relevant sentences, which precludes close reading of much of the material. As well as meeting a practical need for automated analysis, there may also be a methodological preference for data-driven approaches that minimize prior assumptions about the form and content of the textual material, and thus provide less biased, or at least alternatively biased, views on the data which can challenge and extend current theory.

Scientists have been using data-driven approaches for centuries. That is to say, they have at times developed theory based on masses of observations, rather than through hypothesis-driven experimentation. Take, for example, the astronomical Rudolphine Tables which Kepler used to derive laws of planetary motion, and Darwin's extensive and detailed descriptions of animal and plant life that led to the theory of evolution. In more recent times, "discovery science" – now based on automatically detected patterning in vast electronic data sets – has come to be a norm in some branches of science such as astronomy, genomics, and neuroscience. A leading neuroscientist, when considering the relative status and merits of discovery science and hypothesis-driven research, wrote with great candor: "If we were to be truthful, we would admit that often hypotheses are retrofitted to our accidental observations. Moreover, these un-hypothesized observations are, in many cases, more exciting and revealing that anything we would have ever posited. They spur us in directions that no pure thought process ever would. If we are very lucky, they lead us to revise many of the hypotheses previously conceived" (Akil 2003, p. 4).

Importantly, it is apparent that in these statements data-driven discovery science is not seen as an alternative to hypothesis-driven research but rather as a stage prior to the formulation and testing of hypotheses. Indeed, the choice of exclusive alternatives that is implied in the

distinction between “data-driven” and “hypothesis-driven” is unnatural and unhelpful; likewise the parallel distinction between “deductive” and “inductive” research methods, and the “top-down versus bottom-up” metaphor where theory is placed above data. Working deductively, researchers start with a theory that suggests hypotheses to be tested against data generated from an experiment. Working inductively, researchers start with data about that part of the world (or universe) in which they are interested, before detecting patterns and regularities, the explanations for which lead to the formulation of new hypotheses and theory. Of course, in practice, it is not possible to begin selecting data, let alone interpreting it, without some input from prior theory either in the mind of the researcher, or manifested in the computer program used to analyze the data. However, when working more inductively, in an explorative mode, then a researcher is perhaps more likely to make observations that challenge and extend existing theory. There is of course also the risk of wandering fruitlessly with no clear objective through masses of data on a “fishing trip” (a term often used to disparage data-driven research).

Thus, in the context of automatic text analysis, I am interested in what have been termed “discovery tools” (Shneiderman 2001): These are tools that combine data mining and visualization techniques so that researchers can explore alternative views on large data sets. There is always a human in the loop to direct the analysis and, of course, to interpret results and theorize. This resonates with the point that automatic content analysis methods are best for “amplifying and augmenting careful reading and thoughtful analysis” (Grimmer and Stewart 2013, 268). Envisage a researcher faced with a large volume of textual material with which they are not very familiar. Discovery tools should elucidate salient and idiosyncratic linguistic forms in the corpus that could inform decisions about the previously mentioned steps (i)-(iv), as follows.

(i) Selecting non-textual phenomena to investigate. Of course, the researcher likely has some phenomena in mind already when choosing to analyze a particular text corpus, but discovery tools may draw attention to linguistic forms that prompt interest in other phenomena too.

(ii) Determining an appropriate conceptual framework. Again, the researcher will probably have a conceptual framework in mind at the outset, but this may be confirmed or revised by reading the corpus with discovery tools, especially with regards to (iii).

(iii) Establishing a mapping between concepts and linguistic forms that can be counted, i.e. developing a coding scheme. The crux of automatic text analysis is the counting of linguistic forms – words, classes of words, phrases, grammatical structures, etc. – that map to the concepts and non-linguistic phenomena that are pertinent to a particular investigation. If this mapping is established prior to analyzing the corpus then there is a risk of the "seek and ye shall find" problem. That is, if my coding scheme, or my intuition, says that concept X manifests in linguistic forms A, B, and C, then I will never know if there are other ways that X commonly manifests in the corpus, nor if there are other related phenomena and concepts in the corpus that should be considered. A data-driven approach would first identify salient linguistic forms in the corpus. Then some of these forms would be mapped to phenomena and concepts through manual interpretation by confirming or modifying an existing conceptual framework, or by developing one from scratch. This gives a more rigorous and comprehensive mapping between the conceptual framework and the corpus at hand. Of course, this depends on relevant phenomena manifesting in sufficiently frequent and regular linguistic forms such that the forms can be identified with automatic techniques. These techniques can only access the text in contrast with a human coder who can, for better or worse, draw on context and general world knowledge. This places a fundamental limit on the applicability of purely data-driven text analysis techniques.

(iv) Identifying interesting and significant statistical results in the frequency counts of linguistic forms. Once frequency counts of linguistic forms have been generated then the research is akin to other sciences that work with numeric data. So I will not spend time considering this phase here, except to note that the research may continue in either a data-driven or hypothesis-driven manner, and may return to iterate through the earlier steps in the light of the results that are generated.

Given the great diversity of textual material related to climate change, then techniques to elucidate salient and idiosyncratic linguistic forms in a corpus should be applicable and reliable across different languages, text genres and domains. This means that the techniques should have a minimal reliance on prior linguistic descriptions that some techniques depend upon, such as lexicons (vocabularies) and grammars. Such resources are not available for most languages, text genres, and domains, and they are costly to make on a case by case basis. Furthermore, these resources can introduce implicit biases into research. Given that linguists cannot even agree on apparently straightforward descriptions such as parts of speech (to whit multiple part-of-speech tagging schemes, each delimiting a different number of parts of speech), let alone a theory of semantics, then the prospect for computationally tractable linguistic descriptions that are generally applicable seems remote, if not a de facto impossibility. In other words, it seems wise, as far as possible, to let each corpus speak for itself – or "trust the text" (Sinclair 2004) – so that the idiosyncrasies of different domains, text genres, and languages are not lost under generic schemes. To sum up, there are practical and methodological reasons to prefer text analysis techniques that do not depend upon previously created linguistic resources which encode linguistic descriptions and assumptions that cannot be generally applicable.

A brief review of automatic text analysis techniques that have been used to investigate climate change discourse

Broadly, there are two kinds of ways in which automated techniques have been used to investigate climate change discourse, although they are not mutually exclusive: (i) automated techniques have been used to code texts on a greater scale than is possible manually, e.g. by classifying each text in terms of sub-topics or frames; (ii) automated techniques have been used under the rubric of corpus-based discourse analysis to identify and interpret distinctive linguistic patterning.

For automatic coding, typically the text is taken as the principle unit of analysis and each text is treated as a bag of words, i.e. at solely the lexical level such that information about word order and higher-level linguistic structures are ignored. This is commonly the basis for text classification, text clustering, and text scaling; see Grimmer and Stewart (2013) for an in-depth description and discussion of these techniques. Classification techniques assume that classes are known in advance. Often a set of manually classified exemplar texts is used to train an automatic classifier; however this does not seem to have been used for climate change texts. Otherwise, classification proceeds on the basis of combinations of words to define classes of texts; this is similar to the earlier mentioned use of queries to retrieve texts when investigating news media and social media coverage. The mapping between words and classes of texts may be specified a priori or may be developed inductively.

In a simple case, hashtags were used to identify and count tweets associated with different aspects of climate change discourse (Kirilenko and Stepchenkova 2014); in other work, different themes induced from a set of hashtags were used to categorize tweets (Pearce et al. 2014). Six kinds of climate change "narrative" were induced from the transcripts of TV news broadcasts and mapped to detailed combinations of words that define each kind of narrative and enable

automated search for them (Mayer 2012). For example, part of the query for the "climate tragedy" narrative states that either the term "global warming" or "climate change" should occur in the proximity of a word such as "Katrina", "drought", "tornado", "flood", etc.; another part of the query seeks to exclude stories denying climate change by stating that stories should not include words such as "hoax", "alarmist", etc. The categorization was based on "extensive reading of media coverage, as well as conversations with journalists and observers of the media" (5) and the combinations of words were developed by testing "a variety of keyword search strings to find a formula that most consistently identified stories that fit each narrative" (41). A similar approach was taken to analyze climate change frames in Twitter (Jang and Hart 2015), although the combinations of words to identify each frame were somewhat simpler, e.g. a climate change tweet was considered to be an instance of the "hoax frame" if it contained "hoax", "lie" or "fraud". The relevant words were selected from the most frequent words for the top 500 tweets (ranked by the amount they were retweeted).

In the studies mentioned above, establishing the mappings between words and classes was somewhat ad hoc in each case. More systematic analyses may be achieved by using unsupervised techniques for scaling and clustering, such as topic modelling, which are data-driven in that scales and clusters are induced automatically from the texts (which are typically treated as bags of words). Then, insights are gained both from where documents are placed in clusters and on scales, and from inspection of the words that are most discriminatory. Tvinnereim and Fløttum (2015) used structural topic modelling to induce four topics from 2,115 open-ended survey responses in order to characterize public opinion about climate change, and to identify the most representative responses of each topic for close reading. The advantage of unsupervised techniques in general is that they obviate the cost, and possible biases, of manually labelling

some texts as training examples which must be done for supervised techniques. A disadvantage is that sometimes the techniques must be tuned, i.e. to set parameters for the best results, which may mean that many differing outputs must be validated or discounted by the researcher.

It has been established that the processing of texts as bags of words is "sufficient to infer substantively interesting properties of texts (Hopkins and King 2010)" (Grimmer and Stewart 2013, 273). However it is important to recognize that multiple linguistic structures – including many that arise from the sequencing of words in terms, phrases, grammatical constructions, etc. – contribute to the meaning of a text. When treating a text as a bag of words there is a risk of losing most of the meaning arising from these structures so it is important to consider what other aspects of meaning can be elucidated using data-driven techniques. Thus I now turn to corpus-based discourse analysis, i.e. the second kind of way in which automated text analysis techniques have been used to investigate climate change discourse.

Corpus linguistics has well established techniques for exploring linguistic patterning in large text corpora in a data-driven manner whilst ensuring that at least some information about word sequences is kept. Common techniques include frequency lists (giving the frequency of each word), keywords (words that are unusually frequent in a particular corpus), concordances (a set of text fragments around a word of interest), word clusters / n-grams (frequent word sequences including the word of interest) and collocations (words tending to occur with a word of interest); for an introduction to these techniques, see McEnery and Hardie (2012). In recent years there has been considerable interest in applying these techniques under the rubric of corpus-based discourse analysis in order to "uncover linguistic patterns which can enable us to make sense of the ways that language is used in the construction of *discourses* (or ways of constructing reality)" (Baker 2006, 1): A crucial point is that much linguistic patterning would not be apparent to the

researcher without the use of automated techniques. The approach is exemplified in the work of Baker et al. (2008) who used collocation and concordance analysis to identify common categories of media representations (related to refugees, asylum seekers, immigrants, and migrants) and representative texts for qualitative analysis from a 140 million word corpus of newspaper articles. Similar approaches, variously using frequency lists, word clusters, collocates, and concordances, have been used to investigate climate discourses, e.g. concordances of word clusters to analyze "carbon compounds" in web RSS feeds (Koteyko 2010); collocates of frequent words to get a view of the framing of climate change in newspapers (Grundmann and Krishnamurthy 2010); the collocates of environmental terms to determine their use in positive, negative, and neutral contexts (Wild et al. 2013); and, a temporal and geographic comparison of keywords and word clusters in newspapers to identify the main claim makers and the relative visibility of acceptors and sceptics (Grundmann and Scott 2014).

In general, frequent words, keywords, and word clusters can be seen as highlighting salient concepts in discourses. However, these techniques can only be a starting point for a researcher. A lack of information about the co-text around keywords and word clusters restricts the extent to which they can be interpreted without the close reading of concordances. Increasingly, corpora are too large for close reading of enough relevant concordances, so automated techniques are needed to condense information about co-texts, i.e. about the sequences of words around a word of interest. Collocation data gives a view on which words tend to co-occur with the word of interest and hence can be seen as giving insights into the meanings typically associated with it. However, collocation data is typically presented as a large grid of statistics for one keyword and it would be desirable to have a simpler picture that is more intuitive to interpret. Also, whilst collocation data describes a relationship between two lexical items it does not capture further

sequential information. A step towards considering the sequencing of words when analyzing frames has been taken by researchers using text visualization techniques to explore word associations around a term of interest (Baumer et al. 2013; Diakopoulos, Zhang, and Salway 2013; Diakopoulos et al. 2014; 2015). Towards a similar aim, the method for local grammar induction described later in this chapter is intended to provide a condensed view of the patterning in words sequences around a word of interest. First though, the next section presents a case study to reflect on the use of some established corpus linguistic techniques as part of a data-driven approach.

Case Study I: Inducing categories of future representations from climate change blogs

Fløttum et al. (2014) investigated representations of the future in climate change discourse in the blogosphere, using a frequency list, word clusters, sorted concordances, and also simple template patterns. At the outset of this work the interest in representations of the future was established, but little was known and nothing assumed about how such representations might be manifested in the corpus to be analyzed. The NTAP corpus comprises about 3000 English language blogs (1.4 million blog posts) related to climate change issues from 2000-2012 (Salway, Hofland, and Touileb 2013): From this, all 330,000 sentences containing either "climate change" or "global warming" were selected for analysis.

Our first step was to generate a frequency list and inspect the 1500 most frequent words in order to identify those which we judged might form part of representations of the future, e.g. the word "future" itself, and others including "threat", "risk", "danger", and "opportunity". Limiting ourselves to the top 1500 words was an arbitrary but expedient decision, in keeping with the idea of looking for the most salient linguistic forms. The 11 selected words had some 30,000

instances in total, so there was still too much data for it to be feasible to read and analyze all of the relevant co-texts. Instead, word clusters and sorted concordances were generated in order to elucidate patterning around the selected words. Frequency ordered lists of word clusters showed the most common sequences containing each word of interest, e.g. "future of" and "future for". Sorted concordances presented text snippets around the words of interest, sorted alphabetically according to the surrounding words: They were useful to reveal patterns such as "a WORD future", where WORD can stand for any word, see Figure 9.1.

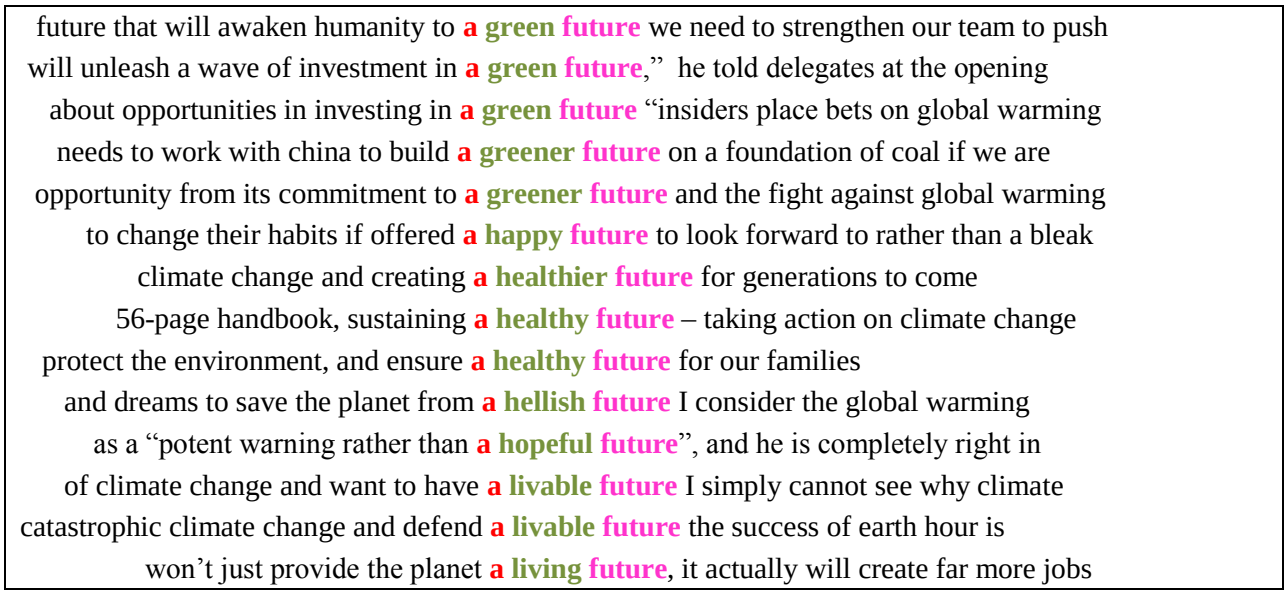

Figure 9.1. Part of a sorted concordance for the word "future" which reveals the pattern "a WORD future"; similar to that used by Fløttum et al. (2014).

Working in this way, 42 patterns related to representations of the future were identified – three of these patterns are shown in Table 9.1. Each pattern contains a filler, i.e. 'WORD'; in some patterns there is alternation which is shown with the '|' symbol and optionality shown with brackets. For each selected pattern a program counted the number of instances of the pattern, the number of different fillers appearing in the pattern, and the frequency of each filler; the table

shows the top five fillers. This data provided a condensed view of the co-texts around the 11 selected words that reveals common ways in which the words are used in the corpus.

Table 9.1. Three of the patterns derived from manual inspection of word clusters and sorted concordances by Fløttum et al. (2014).

Pattern	Unique fillers	Total instances	Number of instances for the five most frequent fillers
a|an WORD future	97	239	sustainable (34); low-carbon (19); better (15); uncertain (12); greener (7)
risk(s)|danger(s)| threat(s) facing WORD	30	142	the (43); our (26); humanity (25); mankind (10); humankind (5)
opportunit(y|ies) to WORD	325	843	make (39); address (18); put (16); build (16); take (15)

Inspection of this data for all 42 patterns, along with the close reading of co-texts for certain pattern-filler combinations, led to the identification of nine meaning categories to characterize the different future representations in the corpus: (1) sustainability, (2) value-laden positive, (3) value-laden negative, (4) temporal, (5) future for people/human beings and future of humanity/planet, (6) future for and of regions/countries, (7) future for nature/environment, (8) future for business/industry/economy, (9) future for security. This categorization provided the basis for further investigations, e.g. to test the hypotheses that "accepting" climate change blogs would be more concerned with the future than "sceptical" blogs, and that, when they did address the future then the "sceptical" blogs would typically be responding to the discourses of the "accepting" blogs (Salway, Fløttum, and Elgesem 2015). The established mapping between the categories and common ways in which they are realized – i.e. certain pattern-filler combinations – meant that it was relatively straightforward to make quantitative comparisons between the two corpora, and to identify samples of future representations in each sub-corpus for close reading.

This work exemplifies the idea of using data-driven techniques to provide a manageable view over a large text corpus, and, in concert with manual interpretation of the results and close reading of samples, to assist in establishing a conceptual framework and a mapping between concepts and countable linguistic forms. It should be noted that, in this example, the mapping between categories and their textual realizations is not comprehensive – rather, as the result of a frequency-led analysis, we expect that it captures the most common textual realizations. Furthermore, we cannot guarantee that every instance of a certain linguistic form (pattern-filler combinations in this case) is being used to convey the same meaning – we assume that most of them are, based on the close reading of some examples. The method that identified salient patterns around frequent words and provided a condensed view of their co-texts (e.g. the examples in Table 9.1) relied on manual analysis of lists of word clusters and sorted concordances, which was somewhat ad hoc and time consuming. In the following section I present and discuss a recently developed technique that generates similar kinds of data about the co-texts around selected words in a wholly automated way.

Local grammar induction: a new technique to induce salient information structures for data-driven content analysis

This technique is intended to induce salient information structures from unannotated corpora as a discovery tool for humanities and social science research. It preserves information about word order in text so that more linguistic structure, and hence meaning, is elucidated than with bag-of-words approaches. Since it does not rely on any linguistic resources, other than a corpus, it should be portable and relatively free of biases. Preliminary results suggest that the technique provides a condensed view of selected words' co-texts which may be useful when analyzing topics, framing, rhetorical strategies, and relations between entities (see case studies II and III).

The technique, described in detail in Salway and Touileb (2014), is based on ADIOS (Solan et al. 2005) which is an unsupervised grammar induction algorithm that induces hierarchical structures from sequential data, e.g. words in the sentences of a corpus. Using statistical information, it identifies the most significant word sequences – referred to here as H(orizontal)-groups, and equivalence classes – V(ertical)-groups. H-groups and V-groups may be nested within others. For presentation, the sequential items in an H-group are separated by whitespace and the alternative items in a V-group are separated by the '|' symbol, such that the output reads like a simple regular expression. For example, take "((climate change)|(global warming)) is caused by". Here the top level H-group comprises a sequence starting with a V-group, which itself contains two alternative H-groups, and then three words. Statistical evidence from a corpus has led the technique to recognize that "climate change" and "global warming" are common word sequences, and that they both often precede "is caused by".

The ADIOS algorithm, like other grammar induction algorithms, 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 our work to modify and apply the ADIOS algorithm for data-driven text analysis. One important modification was to focus the algorithm on text snippets around one single key term of interest at a time, rather than processing all the sentences in a corpus. This change was influenced by the theory of local grammar (Gross 1997), i.e. the idea that language is best described with word classes and combinations that are specific to the local contexts of individual words.

Figure 9.2 shows 10 of the 671 top level H-groups that were induced by Salway and Touileb (2014) from the NTAP corpus (the corpus is described in case study I); the technique was applied separately to 17 key terms ("climate change", "sea levels", "carbon tax", etc.) which ranged in frequency from 1,000s to 100,000s. Each of the 10 top-level H-groups in Figure 9.2 contributes something to an overall impression of what is commonly written in the co-texts of key terms in the corpus. H-groups 1, 2, and 3 each capture alternatives for core domain terminology, e.g. H-group 2 highlights the alternatives "anthropogenic global warming", "manmade global warming" and "man made global warming". Likewise, there are alternatives in 4, but these seem to reflect different strengths of feeling about the need to deal with climate change, i.e. "combat", "minimize" or "tackle". H-group 5 captures four different ways to express that something is caused by climate change or global warming. The next two H-groups are included to show that not all of the output is meaningful: It seems that 6 is incomplete, and 7 has formed a structure with no basis in language. The remaining examples (8-10) include some errors, but nevertheless give a summary impression of common statements, e.g. "pollution blamed for global warming", "should regulate greenhouse gases", and "climate models suggest that", and some of their alternative phrasings in the corpus.

Of course, the examples presented here are in no way surprising to anybody with even a passing knowledge of climate change discourse. However, the point is that an automatic technique was able to generate computationally tractable structures that relate to important content in the corpus. The following two case studies will reflect on how such structures were used to investigate climate change discourse. First I will reflect briefly on the general applicability of the technique.

1. ((carbon|(greenhouse gas)|co2) emissions)
2. ((anthropogenic|manmade|(man made)) global_warming)
3. ((source|emitter|emitters|producers) of greenhouse_gases)
4. ((to (combat|minimize|tackle)) climate change)
5. (((due to)|(caused by)) ((climate change)|(global warming)))
6. ((of global warming) (was|are|is))
7. (in (order|(the (atmosphere|recessions))))
8. (((greenhouse gases)|emissions|gases|(carbon emissions)|pollution) blamed ((for|to) global_warming))
9. ((would|should|to|must) (control|reduce|regulate|regulating|release) greenhouse_gases)
10. (((((global|some|sophisticated|complex|the) climate models)|climate models) (project|suggest|predict)) that)

Figure 9.2. 10 of the top-level H-groups that were induced automatically from the NTAP corpus by Salway and Touileb (2014).

Because the induction process is unsupervised and does not rely on any linguistic resources then it should port effectively and cheaply to different language and domains; an exception may be languages with free word order. That said, because the induction process exploits partially overlapping word sequences around key terms, it should be expected to 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, and especially more stylized varieties of language. Like other unsupervised techniques, the method is sensitive to small changes in the input data, e.g. the order in which snippets are presented. This means that in practice it should be run many times and something done to amalgamate the results from the different runs. It should also be noted that the technique is computationally intensive – it took several days to process about 100,000 snippets for one key term. Of course, the speed of computing hardware continues to improve at a fast rate, but the fact that the ADIOS algorithm does not seem to be conducive to parallelization means that at least some big data technologies

cannot be exploited. The most important caveat is that we are still working to understand more precisely what information structures the technique does and does not capture. Preliminary results are presented and discussed in the following case studies.

Case Study II: Using unusually frequent information structures to highlight distinctive content

The 671 induced H-groups mentioned in the previous section were examined by Touileb and Salway (2014) in order to assess how well they highlighted distinctive content within individual climate change blogs; selected results are presented and discussed here. Table 9.2 presents the top 10 H-groups for one of the three blogs that were focused on – Chimalaya – ranked according to the RRF metric (Edmundson and Wyllys 1961). RRF (ratio of relative frequencies) is a simple measure of keyness that reflects how much more (or less) something appears in corpus A compared to corpus B, whilst factoring in the size of the corpora; in this case, the comparison was between each blog in turn and the other blogs. As before, each H-group is presented with brackets and '|'s; additionally, a breakdown of the frequencies of the various forms is given. For example, H-group 2 occurs 1172 times in total – 1061 times as "developing countries" and 111 times as "poor countries".

The idea was that focusing on unusually frequent H-groups in each blog would elucidate the distinctive characteristics of climate change discourse in each blog. Consider how the top ten H-groups for the Chimalaya blog highlight characteristics that could be useful for further content analysis. Note, Chimalaya is a blog primarily addressing climate politics issues for the Himalaya region; in the corpus there are 3782 posts from this blog, totaling about 3.1million words.

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

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

3. (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)*

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

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

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

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

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

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

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

Figure 9.3. The top 10 H-groups by RRF for the climate change blog Chimalaya, generated by Touileb and Salway (2014).

Many of the H-groups reflect what is already known about the general topical content of the blog, but they also provide a finer-grained view on how the different topics are expressed. The H-groups 1, 3, 5, and 9 indicate Chimalaya's focus on the impacts/effects of climate change, rather than its causes: 3 and 9 include both "causes" and "effects" but the frequencies of the different forms show that this blog is much more concerned with the effects. The blog's interests in addressing climate change are highlighted by 7 and 8, with frequent mentions of meetings in 4. Its focus on the kinds of countries that comprise the Himalaya region is indicated by 2.

H-group 2 "((developing|poor) countries)" also highlights an example of framing: The strong preference for the form "developing countries" (f=1061) compared with "poor countries" (f=111) indicates a choice to frame these countries in a positive way. H-group 8 "(to

(combat|minimize|tackle)) climate change)" is also interesting for framing analysis. It 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 H-group 7.

See Touileb and Salway (2014) for more discussion and comparison of the H-groups in Chimalaya and other blogs, and for a comparison of H-groups with keywords, n-grams and collocation data. Here, for reasons of space, I will mention just one further interesting example of an unusually frequent H-group from a different blog – "((you|we) (can|should))" from Its Getting Hot In Here. This suggests something about the rhetorical strategies used in the blog. 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. 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.

Overall, it seems that skimming a list of unusually frequent H-groups can give a researcher a useful impression of the content of a corpus, and suggest lines of inquiry, e.g. for analyzing topics, framing, and rhetorical strategies.

Case Study III: Inducing structures to analyze international relations and climate positions

As noted previously, the extent to which the local grammar induction technique induces structuring depends in part on how stylized and repetitive the language in the corpus is. When the method was applied to a corpus comprising records of international climate negotiations, which are expected to be rather stylized in form and repetitive in content, then the resulting structures were sufficient to support simple information extraction and the subsequent analysis of country relations and country positions (Salway, Touileb, and Tvinnereim 2014). Structures were induced from a corpus comprising all texts in the Earth Negotiation Bulletin (ENB) volume 12 for the period 1995-2013. All 32,288 sentences mentioning one or more countries were selected, and every mention of a country, or a list of countries, was replaced with the token 'COUNTRY'. This replacement (that used knowledge outwith the corpus) was done to make patterning around mentions of countries more explicit. Seven of the resulting 53 H-groups are shown in Figure 9.4.

1. (COUNTRY ((supported|opposed) by) COUNTRY)

2. (COUNTRY (said|noted|recommended|explained|responded|stressed|questioned| addressed|reiterated|reported|urged|amended|invited…)) ***the V-group contains 51 words*

3. (COUNTRY ((clarified|urged|reported) that)

4. (COUNTRY ((presented|demanded|outlined|favored (the|a))

5. (COUNTRY expressed ((disappointment|concern) that)|((support|appreciation) for)|((readiness|willingness) to)|(satisfaction (with the) (outcome|reconstitution|functioning|work) (of the)))

6. (COUNTRY called (((for |on) (parties|(developed countries)) to)|((for a) (cautious|three phased|common|phased|bottom up|budget|global) approach to)|(for an (overview|elaboration|analysis|evaluation|examination) of)))

7. (COUNTRY highlighted ((the (need|basis) for)|(the (benefits|possibility|establishment) of)|(the (consideration|impact|impacts) of)|(the (use|involvement) of)|((the need to) (err|focus) on)|(the (role|importance) (of the))))

Figure 9.4. Seven of the H-groups automatically induced from Earth Negotiations Bulletin texts by Salway, Touileb, and Tvinnereim (2014).

H-group 1 “(COUNTRY ((supported|opposed) by) COUNTRY)” was used as a regular expression to extract instances where relations between countries were recorded. This gave 1145 instances of support, and 592 of opposition, often involving multiple countries; recall that ‘COUNTRY’ may stand for a list of countries. A count was made for each pair of countries in support and opposition, with a distinction made between ‘C1 supported by C2’ and ‘C2 supported by C1’, and then a scatterplot was made from these counts in order to visualize relations between countries. This highlighted instances of both strong support, e.g. Canada for the United States, and strong opposition, e.g. the European Union to the Group of 77.

Then H-groups 2, 3, and 4 were combined into a regular expression to extract instances of the statements made by countries. For each country a file was made with the text following every instance of “COUNTRY said | noted | recommended | (etc.)”, until the end of the sentence. The country files were then subject to text scaling using the Wordfish tool (Slapin and Proksch 2008): This places texts on a single induced scale which is intended to mirror political positions. For the 40 countries with the most statements, the values of the parameter indicating country position on the induced scale ranged in ascending order from Austria (-2.38) via Belgium, Germany, the UK, … , New Zealand to Japan (-.62) and then on to Papua New Guinea (-.26), Tuvalu, Peru, … , Iran, Bolivia, Barbados, India, and Algeria (1.44).This result – with a large gap between the countries up to Japan and the countries from Papua New Guinea onwards – confirms what is known to be the main cleavage in international climate negotiations between developed and developing countries. Such comparison of countries would not have been possible by scaling the original ENB texts because they are organized by date, not country, so there was value in being able to use the induced structures as the basis for extracting each country’s statements.

H-groups 1-4 all have a relatively shallow structure. In order to induce further structure we created new input files, one for each of the most frequent speech acts from 2-4, e.g. one file comprised all the sentences containing "COUNTRY expressed". The resulting H-groups, including 5-7 in Figure 9.4, do indeed have richer structures and show in a more nuanced way how countries' positions are reported in the ENB texts. As such they could perhaps be used to extract more specific information about countries positions, but this is not something we have tried yet.

Closing remarks

The increasing volume and variety of textual material related to climate change means that it is necessary to use data-driven text analysis techniques as discovery tools in order to elucidate and interpret linguistic patterning. Knowledge of salient linguistic forms is crucial for establishing computationally tractable conceptual frameworks so that the textual realizations of relevant phenomena and concepts can be coded and counted automatically. Taking a data-driven approach to analyzing a text corpus alleviates the problem of "seek and ye shall find" but success depends on the extent to which relevant concepts are realized in regular and repeating ways in texts, as well as on automated text analysis capabilities. Whilst some problems can be tackled by treating texts as bags of words and analyzing language only at the level of lexis, higher-level linguistic patterning must be identified in order to fully realize the potential of automated techniques.

Whilst this can partly be achieved using established corpus linguistic techniques, there is a need for the development of further text mining and visualization tools in order to provide researchers with manageable views of salient linguistic forms in large corpora. Early results from local grammar induction suggest that some meaningful structures can be induced from an

unannotated text corpus, and that giving a researcher access to these structures is useful in the exploratory phase of an investigation. More work is needed to understand what can and cannot be induced from text alone (i.e. without knowledge of context and general domain knowledge), how to manage the sensitivity of inductive techniques to small change in inputs, and how to incorporate data-driven content analysis techniques into discovery tools with informative visualizations and user interaction, and into methods for automatic coding.

References

Akil, H. 2003. "Scientific Strategy in Neuroscience: Discovery Science versus Hypothesis-Driven Research." *Neuroscience Quarterly*, Summer 2003: 4-5.

Auer, M. R., Y. Zhang, and P. Lee. 2014. "The potential of microblogs for the study of public perceptions of climate change." *WIREs Climate Change* 5: 291–296.

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

Baker, P., C. Gabrielatos, M. Khosravinik, M. Krzyżanowski, T. McEnery, and R. Wodak. 2008. "A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press." *Discourse & Society* 19 (3): 273-306.

Baumer, E.P.S., F. Polletta, N. Pierski, C. Celaya, K. Rosenblatt, and G. K. Gay. 2013. "Developing Computational Supports for Frame Reflection." *Proceedings of the iConference*.

Boykoff, M. T. 2008. "The cultural politics of climate change discourse in UK tabloids." *Political Geography* 27 (5): 549-569.

Boykoff, M. T., and J. Timmons Roberts. 2007. "Media Coverage of Climate Change: Current Trends, Strengths, Weaknesses." *Human Development Report 2007/2008*. UNDP.

Diakopoulos, N., A. Zhang, and A. Salway. 2013. "Visual Analytics of Media Frames in Online News and Blogs." *IEEE InfoVis Workshop on Text Visualization*.

Diakopoulos, N., A. Zhang, D. Elgesem, and A. Salway. 2014. "Identifying and Analyzing Moral Evaluation Frames in Climate Change Blog Discourse." *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2014).*

Diakopoulos, N., D. Elgesem, A. Salway, A. Zhang, and K. Hofland. 2015. “Compare Clouds: Visualizing Text Corpora to Compare Media Frames.” *Proceedings of the IUI Workshop on Visual Text Analytics*.

Doulton, H., and K. Brown. 2009. “Ten years to prevent catastrophe?: Discourses of climate change and international development in the UK press.” *Global Environmental Change* 19 (2): 191-202.

Edmundson, H. P., and R. E. Wyllys. 1961. “Automatic Abstracting and Indexing - Survey and Recommendations.” *Communications of the Association for Computer Machinery* 4 (5): 226-234.

Fløttum, K., Ø. Gjerstad, A. M. Gjesdal, N. Koteyko, and A. Salway. 2014. “Representations of the future in English language blogs on climate change.” *Global Environmental Change* 29: 213–222.

Grimmer, J., and B. Stewart. 2013. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis* 21 (3): 267-297.

Gross, M. 1997. “The Construction of Local Grammars.” In *Finite-State Language Processing*, edited by E. Roche, and Y. Schabes, 329-354. Cambridge MA: The MIT Press.

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

Grundmann, R., and M. Scott. 2014. “Disputed climate science in the media: Do countries matter?” *Public Understanding of Science* 23 (2): 220-235.

Harris, Z. 1954. “Distributional Structure.” *Word* 10 (2/3): 146-162.

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

Hopkins, D. J., and G. King. 2010. “A method of automated nonparametric content analysis for social science.” *American Journal of Political Science* 54 (1): 229-247.

Jang, S. M., and P. S. Hart. 2015. “Polarized frames on “climate change” and “global warming” across countries and states: Evidence from Twitter big data.” *Global Environmental Change* 32: 11-17.

Kirilenko, A. P., and S. O. Stepchenkova. 2014. “Public microblogging on climate change: One year of Twitter worldwide.” *Global Environmental Change* 26: 171-182.

Kirilenko, A. P., T. Molodtsova, and S. O. Stepchenkova. 2015. “People as sensors: Mass media and local temperature influence climate change discussion on Twitter.” *Global Environmental Change* 30: 92-100.

Koteyko, N. 2010. “Mining the internet for linguistic and social data: An analysis of ‘carbon compounds’ in Web feeds.” *Discourse & Society* 21 (6): 655-674.

Mayer, F. W. 2012. “Stories of Climate Change: Competing Narratives, the Media, and U.S. Public Opinion 2001-2010.” Joan Shorenstein Center on the Press, Politics and Public Policy Discussion Paper Series #D-72. https://www.hks.harvard.edu/presspol/publications/papers/discussion_papers/d72_mayer.pdf

McEnery, T., and A. Hardie. 2012. *Corpus Linguistics*. Cambridge: Cambridge University Press.

Olteanu, A., C. Castillo, N. Diakopoulos, and K. Aberer. 2015. “Comparing Events Coverage in Online News and Social Media: The Case of Climate Change.” *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2015).*

O'Neill, S., H. P. Williams, T. Kurz, B. Wiersma, and M. Boykoff. 2015. “Dominant frames in legacy and social media coverage of the IPCC Fifth Assessment Report.” *Nature Climate Change* 5: 380-385.

Pearce, W., K. Holmberg, I. Hellsten, and B. Nerlich. 2014. “Climate Change on Twitter: Topics, Communities and Conversations about the 2013 IPCC Working Group 1 Report.” *PLoS ONE* 9 (4): e94785. doi:10.1371/journal.pone.0094785

Salway, A., K. Hofland, and S. Touileb. 2013. “Applying Corpus Techniques to Climate Change Blogs.” *Proceedings of Corpus Linguistics 2013*, Lancaster University.

Salway, A., and S. Touileb. 2014. “Applying Grammar Induction to Text Mining.” *Proceedings of The 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014),* 712-717.

Salway, A., S. Touileb, and E. Tvinnereim. 2014. “Inducing Information Structures for Data-driven Text Analysis.” *Proceedings of ACL Workshop on Language Technologies and Computational Social Science*.

Salway, A., K. Fløttum, and D. Elgesem. 2015. “Representations of the future in "accepting" and "sceptical" climate change blogs.” *Proceedings of Corpus Linguistics 2015,* Lancaster University.

Salway, A., D. Elgesem, K. Hofland, Ø. Reigem, and L.Steskal. 2016. “Topically-focused Blog Corpora for Multiple Languages”. *Proceedings of the 10th Web as Corpus Workshop (WAC-X), ACL 2016.*

Schäfer, M. 2012. “Online communication on climate change and climate politics: a literature review.” *WIREs Climate Change* 3:527-543.

Schäfer, M., and I. Schlichting. 2014. “Media Representations of Climate Change: A Meta-Analysis of the Research Field”. *Environmental Communication* 8 (2):142-160.

Schmidt, A., A. Ivanova, and M. Schäfer. 2013. “Media attention for climate change around the world: A comparative analysis of newspaper coverage in 27 countries”. *Global Environmental Change* 23:1233-1248.

Sharman, A. “Mapping the climate sceptical blogosphere.” *Global Environmental Change* 26: 159-170.

Shneiderman, B. 2001. “Inventing Discovery Tools: Combining Information Visualization with Data Mining.” *Algorithmic Learning Theory* 2225.

Sinclair, J. 2004. *Trust the text: Language, corpus and discourse*. London and New York: Routledge.

Slapin, J., and S.-O. Proksch. 2008. “A Scaling Model for Estimating Time-Series Party Positions from Texts.” *American Journal of Political Science* 52 (3):705-722.

Solan, Z., D. Horn, E. Ruppin, and S. Edelman. 2005. “Unsupervised learning of natural languages.” *Proceedings of the National Academy of Sciences* 102 (33): 11629-11634.

Touileb, S., and A. Salway. 2014. “Constructions: a new unit of analysis for corpus-based discourse analysis.” *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation (PACLIC 28).*

Tvinnereim, E., and K. Fløttum. 2015. “Explaining topic prevalence in answers to open-ended survey questions about climate change.” *Nature Climate Change* 5:744-747.

Wild, K., A. Church, D. McCarthy, and J. Burgess. 2013. “Quantifying lexical usage: vocabulary pertaining to ecosystems and the environment.” *Corpora* 8 (1): 53-79.