View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Repository@Nottingham

Collins, L. (2015) "How can semantic annotation help us to analyse the discourse of climate change in online user comments?" *Linguistik Online* 70(1). doi: 10.13092/lo.70.1743

How can semantic annotation help us to analyse the discourse of climate change in online user comments?

Abstract

User comments in response to newspaper articles published online offer a unique resource for studying online discourse. The number of comments that articles often elicit poses many methodological challenges and analyses of online user comments have inevitably been cursory when limited to a manual content or thematic analysis. Corpus analysis tools can systematically identify features such as keywords in large datasets. This article reports on the semantic annotation feature of the corpus analysis tool Wmatrix which also allows us to identify key semantic domains. Building on this feature, I introduce a novel method of sampling key comments through an examination of user comment threads taken from *The Guardian* website on the topic of climate change.

1 Introduction

The user comments section that follows articles published by journalists online is one format of discussion that is publicly accessible and very popular in the U.K. User comments are enabled on the websites of all major newspapers in the U.K. and in order to contribute users need only complete a free online registration process (it is only *The Times* which requires a paid subscription). Though the timescale in which to make a contribution extends over just a few days and the comments are subject to moderation, the 'thread' is archived and remains publically viewable. Even in the space of a couple of days, articles often attract in excess of

1000 comments and as such, provide a rich resource for the examination of attitudes and opinions around climate change.

The amount of data generated online poses challenges for researchers to gather a more extensive account of such discussions across time, across formats, across media publications, even across individual articles. Previous research looking at online user comments has generally been based on manual content analysis and as such has been limited in the scope with which it can represent online debates (Manosevitch and Walker, 2009; Milioni *et al*., 2012; Coe *et al*., 2014). Such content analysis requires a very close reading of the data in order to construct a coding framework which is very demanding, given the size of the data. In addition, the method warrants multiple 'coders' to generate inter-rater reliability. Milioni *et al*. (2012: 29) describe a pre-coding pilot of separate data and state that they were limited to a sample of the full dataset they wanted to explore, which has implications for what can be extrapolated about the data set as a whole. Gabrielatos *et al*. (2012: 171) assert that the examination of electronic web-based texts requires "the development of techniques which can deal with the complexity that such data expose the analyst to".

Due to the size of the available data when examining user comments online it is natural to consider a computer-assisted analysis. Researchers might adopt a 'corpus-based approach' (Tognini-Bonelli, 2001) whereby they look to validate, refute or refine a pre-conceived idea about the dataset. Alternatively, we can use corpus analysis software to take a corpus-driven (Tognini-Bonelli, 2001: 85) or a 'data-driven' approach whereby: "decisions on which linguistic features are important or should be studied are made on the basis of information extracted from the data itself" (Rayson, 2008: 521). This is not to say that this is an entirely inductive approach, since pre-existing ideas about language categorisation must be present in

the programming of the software that annotates the data. Annotation is, in itself, a process of adding interpretive, linguistic information to data collected as a corpus (Leech, 1997). In this paper I use the Wmatrix corpus analysis tool developed at Lancaster University (http://ucrel.lancs.ac.uk/Wmatrix/), which offers a predetermined framework that can be applied consistently across data types. Its annotation features are explored below.

2 Aims

In this paper I demonstrate how semantic annotation can offer greater insights into online discourse than conventional keyword analysis. It is hypothesised that because semantic analysis of the data incorporates more individual terms it can provide a more comprehensive account of the key themes of the data compared to individual keyword analysis. Furthermore, building on the semantic annotation function of the corpus analytical software tool Wmatrix, I introduce a novel method for identifying key comments within the thread, based on their containing the constituent words of those key semantic categories. This offers a form of sampling that would enable researchers to incorporate more data points into their analysis and examine user comments in more depth. I discuss the implications and limitations of this sampling technique, before considering how researchers might then go on to examine their sample in greater detail.

3 Methods

3.1 Data

The Guardian has enabled readers to make comments online since March 2006 (Hermida and Thurman, 2008). A search was conducted through *The Guardian* website's archive from 2006 until June 2013 for the term 'climate change'. According to the NRS Digital Print and Digital Data survey, *The Guardian* had the largest readership of what were termed the 'Quality

newspapers' (which included *The Daily Telegraph*, *The Times*, *The Independent and The Financial Times*) with 6.4 million visitors each month (http://www.guardian.co.uk/news/datablog/2012/sep/12/digital-newspaper-readerships-national-survey?INTCMP=SRCH). Though the topic is discussed in alternative terms, such as 'global warming', 'climate change' was deemed to be more inclusive and more prevalent terms in the debate (http://guardianlv.com/2014/05/climate-change-a-more-accurate-term-than-global-warming/). A total of 30 752 articles from *The Guardian* were identified through the search term 'climate change'. The search showed that very few comments were made on articles relating to climate change before 2008. This may indicate some delay in the feature being taken up substantially by the readership or may be because there was originally a limit of 50 comments set for each article. In order to test the semantic annotation function of Wmatrix against the largest dataset the articles were sorted according to the greatest number of comments elicited and articles making only a passing reference to, for example, "Chris Huhne, Secretary of State for Energy and Climate Change" were excluded. Thirty-three 'climate change' articles from *The Guardian* website elicited 500+ comments, with the highest being 1679 comments. This demonstrates the depth of information available for conducting a longitudinal, cross-case comparison between articles and between newspapers. The top three ranking articles by number of user comments elicited were identified for analysis. Other researchers might consider alternative criteria such as the date, authorship, or source material in the gathering of their data. The three articles with the highest number of comments taken from *The Guardian* website can be seen in Table 1.

Table 1 Highest ranking articles by number of user comments

Title	Date	Author	Comments
1. "That snow outside is what Global warming looks like" (http://www.guardian.co.uk/commentisfree/2010/dec/20/uk-snow-global-warming?INTCMP=SRCH)	20th December 2010	George Monbiot	1679

2. "The climate denial industry is out to dupe the public. And it's working" (http://www.theguardian.com/commentisfree/cif-green/2009/dec/07/climate-change-denial-industry)	7th December 2009	George Monbiot	1422
3. "Global warming rigged? Here's the e-mail I'd need to see" (http://www.theguardian.com/commentisfree/cif-green/2009/nov/23/global-warming-leaked-email-climate-scientists)	23rd November 2009	George Monbiot	1296

Kehoe & Gee (2012) have examined the relationship between articles and their user comments threads, considering whether the threads can operate as an indicator of the 'aboutness' of the original article. Furthermore, most newspapers have dedicated journalists who will publish material on specific topics and it is no surprise to find that the three articles identified here were written by the same author, George Monbiot. Here, there is potential to consider the role of the journalist and of the public (at least those who contribute to comments threads) in shaping the debate and whether the discussions actually relate to the original post or if they are simply another platform to have a broader discussion. This is something that would require close examination over a period of time and an analysis of a particular online community but will not be explored here.

From these three articles alone, a total of 4397 comments (approximately 484 000 words) was extracted. With the first of these articles, the first comment was submitted 5 minutes after the article was posted and the final comment was submitted nearly three full days afterwards. It would appear that soon after this the comments section was closed by the moderator. Based on the comments sections from the other articles, this seems to be a fairly typical timeframe in which users are given the opportunity to contribute. Some comments were removed by a moderator (95; 80; and 121 respectively), which on *The Guardian* website is identified with a standard message that also incorporates a link to the site's community standards (http://www.guardian.co.uk/community-standards) and FAQs

(http://www.guardian.co.uk/community-faqs), but which remain in the thread to indicate when they were posted and who posted them. We must be aware that this can affect our understanding of the public debate as it exists online, with nearly 10% of the comments made in response to the third article removed. This raises the question of the democratic potential of online user comments as a space for open discourse as traditional journalists grapple with maintaining a certain standard of debate in relation to their publication. There is evidence in the remaining comments that users have to consider the practice of moderation not only in what they go on to write but also where they post it, often referring to other blogs or forums with different moderation policies.

3.2 Keyness

In adopting a corpus- or data-driven approach to determine what is of interest or significant within the data we attempt to capture its 'aboutness' as a matter of frequency (Scott, 1999). More commonly, this is referred to as 'keyness', which Baker *et al.* (2008: 278) define as "the statistically significantly higher frequency of particular words or clusters in the corpus under analysis in comparison with another corpus". In this respect, what is 'key' to a text is determined by recurrent themes, ideas or concepts. This is in contrast to the way in which the term was used in the work of Williams (1983), where 'keywords' were identified as those words with some social, cultural or political significance. In its quantitative sense, 'keyness' is based on relative frequency and as such necessitates a comparison with a 'normal' frequency of words across a stretch of discourse. In corpus analysis this is determined by a reference corpus, which is traditionally a larger dataset suitably matched to the type of discourse under examination. Scott and Tribble (2006) were able to demonstrate that the process of determining keywords in a single dataset is fairly robust, by comparing *Romeo and Juliet* to various, increasingly obscure reference corpora. However, Culpepper (2009: 35)

maintains that “the closer the relationship between the target corpus and the reference corpus, the more likely the resultant keywords will reflect something specific to the target corpus”. Wmatrix has the British National Corpus (BNC) built in to its software, offering both spoken and written language data as well as a number of subdivisions of this larger corpus determined by the context in which the data was captured (business, education and leisure, for example). The BNC comprises 100 million words used in spoken and written language. The reference corpus used by the Wmatrix tool uses a smaller sample of its written data containing 968 267 words. There is no sub-corpus that specifically comprises computer-mediated discourse however and we must have consideration for discourse features that could be determined by this online format. Since the comments under examination seemed to retain a written style the BNC written sampler was used as the reference corpus.

The default statistical measure for determining keyness in Wmatrix is log-likelihood. There are a number of statistical measures that can be applied to determine keyness, such as Pearson’s chi-squared and Fisher’s Exact Test, however log-likelihood is the preferred measure (for a more in-depth survey, see Rayson, 2003). Log-likelihood can be thought of as a measure of ‘difference’. It is calculated through a contingency table which takes into account the frequency of the word in relation to the total number of words in the corpus and compares those to the corresponding values in a reference corpus (for a full account of the calculation see Rayson, 2008). Log-likelihood is presented as a number, the value of which indicates the ‘difference’ to the reference corpora in that a value of zero indicates a perfect match. A negative value indicates that the word is under-represented in the target corpus and a positive value indicates that the word occurs more often than ‘normal’. Furthermore, the higher the value the more significant the difference, with the following critical values:

- A log-likelihood value of 3.84 represents a p-value of <0.05.

- A log-likelihood value of 6.63 represents a p-value of <0.01.
- A log-likelihood value of 10.83 represents a p-value of <0.001.
- A log-likelihood value of 15.13 represents a p-value of <0.0001.

Keywords are generally presented in a frequency table. The critical values provide some justification for setting a cut-off for the number of keywords to investigate, or the researcher can set a specific p-value, as in WordSmith (Scott, 2007). What has been observed however, is that although in most scientific disciplines a p-value of 0.05 is more than satisfactory, even at a p-value of 0.001 the researcher is left with a great number of words to investigate (Berber Sardinha, 1999). Subsequently, the researcher must rely on alternative means of establishing a cut-off point. Berber Sardinha (1999: 4) suggests extracting a majority, i.e. half the total keywords plus one. This is one of the issues Rayson (2008) cites in support of a semantic category analysis, where there will be fewer items as words are collected in groups. He also argues that this type of grouping would promote low-frequency words that individually might be overlooked, in instances where they belong to a key semantic category. Wilson (1993: 3) remarks upon the limitation of a word-based frequency count in that "people also tend to repeat the same concept within a discourse in somewhat different words through the use of virtual synonyms or the negation of a positive attribute". Thus, if a speaker wanted to testify to the size of something they might use a combination of 'large', 'big' and 'massive', the quantitative effect of which would be lost in a single word frequency table. Though there may be fewer categories, the researcher might still want to consider the constituent words individually. Ultimately, this offers a different kind of keyness and a different representation of the text, as will be seen below.

3.3 Keywords in context

Researchers very rarely comment upon keywords in isolation, but refer to the context in which they occur by looking at concordance lines and considering features such as collocation, which identifies "statements of the habitual or customary places of [a] word" (Firth, 1968: 181). Grundmann and Krishnamurthy (2010) analyse collocations using WordSmith in their international examination of climate change discourse in traditional newspapers. Collocation is referred to here to mean the above-chance frequent co-occurrence of two words within a pre-determined span (i.e. three/four/five words on either side of the word under investigation (Hoey, 1991). They report a contrast between the use of neutral collocates of '[climate] change' in the U.S. press (*Intergovernmental Panel on Climate Change*, *research*, *scientists*, for example) and the collocation of action items in the U.K. press (*tackling*, *combat*, *threat* and *levy*), which is also observed in the German and French press (Grundmann and Krishnamurthy, 2010: 128) (see also similar results using multidimensional scaling in Nerlich *et al.*, 2012). They also report a general trend in the U.S. of discussing climate change at a national level (through the words *state, people, president* and *Bush*), compared to the more international framing of the debate in the U.K. press (Grundmann and Krishnamurthy, 2010: 124). Researchers have also explored 'semantic prosody', as a type of extended collocation that is "spread over a unit of language which potentially goes well beyond the single orthographic word" (Partington, 2004: 132). This does provide more contextual information than singular keywords, as well as an indication of the semantic fields which are associated with keywords. However, this process of analysis still relies on specific terms being used and would not be sensitive to the use of near-synonyms or alternative phraseology. In order to determine the different meanings of homonyms grammatical or semantic annotation is required. Corpus analysis software can distinguish between the noun and verb forms of *stick* for example, through grammatical

annotation. Similarly, with semantic annotation such tools can distinguish the *bark* of a dog from that of a tree.

Koteyko *et al.* (2013) explored user comments on articles to do with climate change extracted from *The Daily Mail* website and manually identified a sub-corpus based on the word 'science', which was identified as a keyword. They were able to examine the role of science and scientists in the climate change debate by looking at references to 'science' in context (primarily through concordance lines). Other researchers such as Bassi (2010) have conducted a manual categorisation of keywords into semantic categories to look at the broader themes around the Kyoto protocol. The Wmatrix semantic tagging system collates all members of the 'word family' (Bauer and Nation, 1993) 'science' (*scientist*, *scientific* etc.) as well as other science-related terms (such as *physics*, *nuclear*, *experiment*) within a single category in a more comprehensive way than Koteyko *et al.* (2013) had done manually. This enables researchers to consider a sub-corpus in isolation or in comparison to other such categories.

3.4 Semantic annotation and key categories

Corpus analysis software such as Wmatrix can annotate data for its grammatical and semantic components. Part-of-speech (POS) tagging assigns each word a grammatical label. From the grammatical annotation of a word and the words around it the software can then separate the semantic meaning of homographs. Wmatrix (http://ucrel.lancs.ac.uk/Wmatrix.html) is a corpus analysis tool that was developed at the University Centre for Computer Corpus Research on Language (UCREL) by Dr Paul Rayson as part of the Reverse Engineering of Requirements to support business process change (REVERE) project (Rayson *et al.*, 2000). The POS-Tagging system built in to Wmatrix is the Constituent Likelihood Automatic Word-

tagging System (CLAWS) and has been continuously developed since the early 1980s (Garside, 1987). This system contains a lexicon of words and multi-word units (e.g. such_as, given_that) as well as a list of suffixes to help identify unknown words. This is an advantage over most forms of topic modelling that rely on single word categorisation and has been shown to be more effective in capturing key themes in the data (Lau *et al.*, 2013). A full exploration of its features, as well as a tutorial in using the software can be found here: http://ucrel.lancs.ac.uk/Wmatrix3.html.

Following the standard POS-tagging, the Wmatrix software also conducts semantic annotation, with its unique built-in UCREL Semantic Analysis System (USAS). This tagging system allocates each word of the data into one of 21 major discursive fields (for the full list see: http://stig.lancs.ac.uk/Wmatrix3/semtags.html). An example of how the data is tagged by Wmatrix is shown in Table 2 which is provided by the Wmatrix webpage (http://ucrel.lancs.ac.uk/annotation.html#POS):

Table 2 CLAWS and USAS tagging

Grammatical Tag (CLAWS)		**Semantic Tag (USAS)**
PPIS1 1st person sing. subjective personal pronoun (I)	*I*	**Z8** Pronouns
VV0 base form of lexical verb (e.g. give, work)	*like*	**E2+** Like
AT1 singular article (e.g. a, an, every)	*a*	**Z5** Grammatical Bin
JJ General adjective	*particular*	**A4.2+** Detailed
NN1 singular common noun (e.g. book, girl)	*shade*	**O4.3** Colour and Colour Patterns
IO of (as preposition)	*of*	**Z5** Grammatical Bin

NN1 singular common noun (e.g. book, girl)	*lipstick*	**B4 C**leaning and Personal Care

This process is automated however the user does have the capacity to review alternative tags to those words where meaning can be ambiguous. The software provides a string of alternative tags in order of descending probability, which is informed by its built-in dictionaries. For example, in the construct 'a single man' the tagging system locates 'single' in the category of 'Quantities'. The system also offers 'Not part of a group', 'Relationship: Asexual' and 'Vehicles and transport on land' (presumably in reference to a single decker bus). In each case the software has identified 'single' as an adjective and offered the most common use of the term, to mean: individual, solitary, lone. This emphasises the importance of continually adding to the templates that inform these tools to provide more accurate accounts of language in use. The disambiguation phase involves seven dimensions: the POS-tag; general likelihood ranking for single-word and template tags; overlapping template resolution; domain of discourse; text-based disambiguation; contextual rules; and local probabilistic disambiguation (Rayson, 2003: 67-68). In this way the USAS-tagger will use information such as the grammatical tag, the frequency of the semantic sense of a word in the reference data (for example, 'green' being referred to more commonly as a colour rather than as being environmentally friendly), the known domain of the surrounding discourse and the premise that a word carries consistent semantic meaning throughout a text in order to allocate the most likely semantic tag.

Other methods of semantic tagging include: LIWC (Linguistic Inquiry and Word Count) (Tausczik and Pennebaker, 2010); WordNet (Fellbaum, 1998); and UKB (Agirre *et al.* 2009). However, LIWC does not carry out word disambiguation by drawing on the context of the

word in the manner described above. WordNet is able to conduct semantic disambiguation but does not categorise closed-class words such as pronouns, prepositions and conjunctions. The Wmatrix CLAWS and USAS have high reported figures for accuracy (96-97% and 91% respectively) (Wilson, 1993; Rayson *et al.*, 2004). Piao *et al.* (2004) demonstrated that the USAS can account for 99.39% of the BNC spoken data and 97.59% of the BNC written data. Other research has shown that the CLAWS and USAS can effectively account for the linguistic features of historical English texts (for example, Archer *et al.*, 2003). Culpepper (2009) however had some problems in analysing individual characters' idiolects because of the change in semantic meaning of particular words from the sixteenth century. We may be critical of the reported 91% accuracy of the USAS however these percentages can be improved with the continued addition of more data and it was found to have a higher accuracy in its application here (see below).

3.5 Applications of the USAS

For the most part, applications of Wmatrix and its USAS have been conducted within the research team at Lancaster University where it was developed. Rayson (http://ucrel.lancs.ac.uk/Wmatrix3.html) demonstrates how the USAS can be used to offer a comparative analysis of the Liberal Democrat and Labour party manifestoes for the 2005 General Election. Other researchers have also applied the semantic tagging system in comparative analyses of two texts, such as Murphy (2006) in the analysis of Shakespearean soliloquies; Kaur (2010) in analysing the themes that characterise Malaysian boys' and girls' creative writing; and Potts and Baker (2012) in comparing the cultural differences between British and American English over time. Cheng and Lam (2013) use Wmatrix's semantic annotation feature to compare Western and Chinese media representations of Hong Kong over time, during and after the transfer of sovereignty. L'Hôte and Lemmens (2009)

effectively demonstrate that semantic tagging offered insights beyond referring to keywords around ideas of ‘newness’ in their analysis of New Labour manifestoes.

The use of Wmatrix across research domains demonstrates its utility as a ready-made analytical tool but we must be critical of accepting the USAS categorisation framework as standard. In most cases, researchers used the semantic tagging as a preliminary step in their analysis before examining the findings in more depth. In the examples given above the researchers do not challenge the boundaries of the USAS categories but it is worth considering how certain readings of the text are privileged by such a system. Culpepper (2009) applied the USAS to the speech of specific characters in *Romeo and Juliet* and was able to report patterns identified by the software that were not intuitive or easy to predict but seemed “well-motivated”. For example, we may not be conscientious of a high use of conjunctions at a numerical level but when this is made known through frequency lists it may validate our reception of the style. This showed that there are insights that can be offered by such computational approaches that might not be apparent to us on first reading but which seem agreeable when supported by numerical data. Culpepper (2009: 55) does however state that the advantages afforded by semantic tagging appear to favour ideational categories of language. This observation refers to Halliday’s (1994) ‘metafunctions’, indicating that the semantic tagging process is less suited to providing insights into textual and interpersonal categories beyond those afforded by standard keyword analysis.

3.6 Identifying ‘key comments’

Each full discussion thread was tagged separately by Wmatrix. The comment ‘signature’, which includes the users ‘name’, an avatar (if chosen), a timestamp for the comment and a tally of recommendations, was removed so that only the content offered by the user was

included. This did include citations of other comments, which creates the potential for a lot of repetition but since it is the user's choice to incorporate that content into their own message, it remains an important aspect of their contribution. User comments taken from *The Guardian* website generally maintained a conventional written structure and as such, the BNC written sampler was used as the reference corpus. This produced the frequency tables identifying keywords and key categories.

Though the USAS tagging system does draw on grammatical and semantic information from the context of the word to achieve semantic disambiguation, returning to observe keywords in context allows us to investigate further aspects of style. In order to view the constituent words in context they were tagged in the original discussion thread. Each semantic category was coded with a different colour so that the distribution of concepts was visible on the page. The words were coded using Microsoft Word's 'Replace' function. This was largely an automatic process in that the software is able to locate each instance of the given word and reformat the font style of each occurrence. However, since the word processor does not have the annotation of Wmatrix it can identify only orthographic forms. As such, it is unable to distinguish between those instances of the word relevant to the semantic category in question and those that are not. This required manual checking but the researcher can simply refer to the annotation of Wmatrix through, for example, concordance lines in order to match the instances of a word that have been allocated a particular semantic tag. Given that Wmatrix already has the function to view a particular word in the context of the entire file and highlights the word in blue, this problem could be overcome if multiple words could be highlighted in one presentation of the file. Wmatrix can of course distinguish occurrences of the word relevant to the semantic category in the way that a word processor is unable to do,

making this another process of the methodology that could potentially be completed automatically by Wmatrix.

Once the tagging was complete I was able to visually observe each occurrence of the words within the top ten categories in their original context. In this way, we can view sequentially how particular topics converge and the density with which they co-occur. There are similarities here with the dispersion plot feature of WordSmith, which gives a visual indication of the degree to which keywords are localised or well-distributed across a text (Scott, 2007). We can view each post as a micro-unit of analysis that has clear parameters. This is not to say that the posts exist in isolation, since they are of course part of a longer cohesive thread. But nor do they adhere to a strict linearity, since posts are often directly aimed at other contributions that appeared some time before in the sequence. Commenters often made a series of points within the parameters of a single post, the inference being that they incorporated all of the dimensions of the discussion that they deemed to be appropriate and relevant at that point in the discussion thread. Thus the comments should not be understood as 'turns' in the way that conversation analysis understands each contribution and the sequence of utterances in this online format does not strictly adhere to the linearity of other types of discourse. This principle requires further investigation but for the purposes of this study each comment was taken as a cohesive unit.

The next phase of this method was to extract a sample. Since semantic annotation was able to identify key semantic categories in the data the sample was to be determined by key themes. I began by locating those comments that engaged with those key themes i.e. that incorporated those key categories. Such a sample is by no means representative of the full treatment of each concept within the discussion. Nor is it representative of the participants of the debate,

of this particular thread in this particular publication, let alone the wider public discourse. What this extraction process did provide was a much smaller sample of comments that allowed us to observe how those key semantic categories operated in relation to one another. In setting the criterion that comments incorporated all ten of the key semantic categories I identified a particular subsection of the discussion which considered the 'full picture'. In such comments, users necessarily offered a much broader perspective on the discourse. This seemed to be a suitable starting point to get an initial sense of the prevalent themes in the debate and how they were seen in relation to one another. The researcher might take any number of those ten key semantic categories (or more) and consider more specific relationships between any number of them: 'science' and 'weather' for example. Unsurprisingly, the comments which included all ten categories were much longer than the average. This process therefore, favours those who make more substantial contributions, as a longer comment is more likely to incorporate more of the key semantic categories. The dual effect is that those more substantial contributions will have a greater effect on what those key semantic categories are, simply by constituting a larger part of the data.

4 Analysis

4.1 Manual correction

The semantic tagging process did require some manual correction. Those words which were reallocated were almost exclusively assigned into the category 'Z99 Unmatched', indicating that the software simply did not recognise them. These words generally fell into three types: compound adjectives; groups of people within the debate characterised by their beliefs; and personal names. The compound adjectives (*factory-produced, low-power, god-shaped, eleven-dimensional, coal-fired, carbon-containing, coal-powered, carbon-neutral, bio-fuelled*) in some cases were reallocated to the category 'W5 Green Issues' but for the most

part belonged in the category ‘O4.1 General appearance and physical properties’. The names for groups of people within the climate debate based on their beliefs (*warmists, denialists, catastrophists, armageddonist, doomsayers*) were generally reallocated into either ‘X2.1 Thought/belief’, ‘W5 Green Issues’ or ‘S5+ Belonging to a group’. The unique usernames used by commenters (*Bluecloud, gourdonboy, jbowers, georgecoldwell, macsporan, lovelock, HypatiaLee*) were generally re-allocated into the category ‘Z1 Personal Names’.

These usernames are characteristic of online discourse in that commenters rarely use their given names. This demonstrates a kind of creativity as part of an online persona but can also be seen as part of the online disinhibition effect (Suler, 2004), where individuals are in many ways less accountable for what they post. Since they are unlikely to have appeared in the reference corpus a keyword analysis is likely to identify these unusual names as significant, even with a low occurrence. This is similar to Scott’s (2007) observation of a text about horse racing wherein the horse’s names would be quite incidental to the story but would gain statistical significance with very few mentions. They are correctly identified by the software as (proper) nouns and thus allocated to the ‘Z’ semantic category, but the software does not recognise which subcategory is appropriate. Personal names – particularly those used online – demonstrate a great degree of creativity and as such, are likely to require some degree of manual tagging in order for their semantic quality to be properly recognised.

Ideally, there would not be any words tagged in this ‘Unmatched’ category but the ever-changing nature of language and creativity of those who use it means it is unlikely that we will be able to account for every word within a text thruogh computer programming. If the researcher is content with the USAS’s reported 91% accuracy it may be sufficient to simply reallocate those words which have remained ‘Unmatched’. The number of corrections in the

'Unmatched' category was minimal in relation to the size of the overall corpus: 1146 of 60931 tags (1.88%) for the first thread, 3138 of 164810 tags (1.90%) for the second thread and 1497 of 81200 (1.84%) tags for the third thread. Despite only accounting for those tags assigned to the 'Unmatched' category, these figures are well within the range reported in the literature. Nevertheless, in the first user comment thread manual correction elevated the significance of the category 'Z3 Other personal names' to appear as one of the top ten categories. This category was elevated from a log-likelihood value of -3.44 (i.e. underused) to + 814.46 and a similar elevation was evident in the other two comment threads. Had manual correction not been applied, the category 'O1.3 Substances and materials: Gas' (which included the words *CO2, air, gas, methane* etc.) would have appeared in the top ten categories for the first thread.

4.2 Keywords and key categories

The top ten keywords for each user comment thread are represented in Table 3.

Table 3 Top ten keywords for each comment thread

	Thread #1	**Thread #2**	**Thread #3**
1.	climate	climate	science
2.	warming	data	climate
3.	global	warming	warming
4.	AGW	science	data
5.	weather	that	scientists
6.	science	global	global
7.	winters	you	emails
8.	is	is	you
9.	you	scientists	CO2
10.	change	CO2	AGW

Since the articles were identified through the search term 'climate change' it tells us little that in the user comments discussion, 'climate', 'change', 'global', 'warming', 'AGW' (Anthropogenic Global Warming), 'weather' and even 'winters' occurred more frequently

than ‘normal’. The high occurrence of ‘science’ was also no surprise since it is central to the debate. ‘Is’ and ‘you’ (which was found in all 3 lists) are used with such frequency and variety that it is difficult to interpret their use, though we may infer a preoccupation with a current state of affairs, of what ‘is’ and what ‘is not’. Similarly, we might infer a prevalent dialogic style, though ‘you’ is often used in its universal and non-specific sense (i.e. ‘one’). Ultimately, three sets of ten words alone offered a very narrow perspective of the defining features of this discussion.

The top ten semantic categories from the user comment threads following manual correction can be seen in Table 4.

Table 4 Top ten key semantic categories from user comments thread

	Thread #1	**Thread #2**	**Thread #3**
1.	W4 Weather	Y1 Science and technology in general	Y1 Science and technology in general
2.	O4.6+ Temperature: Hot/On Fire	X2.2 Knowledge	A5.2+ Evaluation: True
3.	Y1 Science and technology in general	A5.2+ Evaluation: True	X2.2 Knowledge
4.	A5.2+ Evaluation: True	Z3 Other proper names	Z3 Other proper names
5.	Z3 Other proper names	W4 Weather	Z8 Pronouns
6.	O4.6- Temperature: Cold	Z6 Negative	W4 Weather
7.	A3+ Existing	A3+ Existing	A3+ Existing
8.	A2.2 Cause & Effect/Connection	Z8 Pronouns	Z6 Negative
9.	O4.6 Temperature	O1.3 Substances and materials: Gas	O1.3 Substances and materials: Gas
10.	Z6 Negative	O4.6 Temperature	X2.1 Thought, belief

Choosing the top ten categories may seem quite arbitrary but given the high degree of significance at which these categories occurred there is little rationale for establishing a threshold. As indicated above, a log-likelihood value of 15.13 relates to a p-value of 0.001. In the first discussion thread the tenth most significant category had a log-likelihood of 606.07

and the top semantic category a log-likelihood value of 2640.24. There is no question of the significance of the occurrence of words in these categories but it is the researcher's decision to establish parameters of what they take forward to the next stage of analysis.

We can observe some consistency between the keywords and the key semantic categories from the words 'weather', 'climate' and 'winters' which come under the semantic category 'W4 Weather'; and the occurrence of the words 'science' and 'temperature' with their respective semantic categories. In the first thread there were three separate categories concerned with 'temperature'. This reflected both the content of the article, which considered how the presence of snow is a result of the changing weather system and can be attributed to more general changes in the climate; but also a tendency to discuss climate change as a rise or fall in temperature, incorporating the debate about the misnomer 'global warming'. The category of 'Science and technology' was shown to be very significant in all three lists, as was a preoccupation with 'evidence', 'facts' and 'truth' in the category 'A5.2+ Evaluation: True'. The category 'Cause and Effect' suggested that climate change is understood in terms of its potential causes and effects on, for example, the weather. A more deductive approach at this stage might be to consider the relationship between 'science' and 'climate' as discussed in this thread since of 1679 comments, 1467 (87.4%) made at least one reference to 'Weather', 'Temperature' or 'Science'. This would offer a more topical focus that is validated by the data but is not pursued here.

In the same way that 'is' was a keyword, the category of 'Existing' was significant in all three threads. Similarly, the category, 'Negative', which included terms of negation such as 'not', 'none', and 'neither' was significant in all three threads, reflecting a tendency for the discussions to be around what 'is' and what 'is not'. Koteyko *et al.* (2013) also found that

'not' was significant keyword in their analysis of user comments taken from *the Daily Mail*. In both cases this was often in reference to science and scientists, demonstrating that questions as to the legitimacy of climate science and the practices of climate scientists are key in the climate change debate. This also supported the view that the debate is characterised by polarised opinions. The tendency to write in terms of what 'is' and what 'is not' reflected a claim followed by a counter-claim interaction, with little indication of mediation. The frequent use of personal names (apparent in the prominence of the category 'Z3 Other proper names' in all three threads) also attested to a degree of user interaction.

For the most part, keywords were incorporated into one of the top semantic categories for each thread. Key categories offer a slightly different perspective on what characterised the discussion as a whole but still indicated which individual terms within those concepts were of significance. Semantic tagging allows us to speculate about broader patterns in the data but at this level we do not get accurate or specific details about how these key concepts operate in context. To refer to concordance lines would warrant a great deal of work, since there are multiple word units to consider within each category. What is proposed here is that the constituent words of the kay categories are tagged on the original discussion thread as described above in order to observe the dispersion of those key concepts as well as identify key comments in the thread.

4.3 Key comments

In the first discussion thread 17 of the 1679 comments (approximately 5 300 words out of an original 163 000) incorporated all ten of the key semantic categories. To give an example of the specificity of this approach, if we were to extract comments that exhibited any nine of the ten categories we would have a sample of 64 comments (16 500 words), which is nearly four

times as many comments and three times as many words. Again, the individual researcher can adjust the parameters based on the size of the sample they are looking to extract but with a view to making multiple comparisons it is preferable to extract a smaller yet data-rich sample. The number of comments and word count for the data extracted at the parameters of incorporating eight or more of the top ten semantic categories for each thread are shown in Table 5.

Table 5 Number of comments and word count for key comment extraction

	Original texts		**8 or more categories**		**9 or more categories**		**All 10 categories**	
Thread	**Comments**	**Words**	**Comments**	**Words**	**Comments**	**Words**	**Comments**	**Words**
#1	1679	163 180	159 (9.47%)	36 103 (22.12%)	64 (3.81%)	16 451 (10.08%)	17 (1.01%)	5 264 (3.23%)
#2	1422	182 636	154 (10.83%)	56 922 (31.17%)	56 (3.94%)	25 593 (14.01%)	20 (1.41%)	11 352 (6.22%)
#3	1296	138 304	172 (13.27%)	45 718 (33.06%)	68 (5.25%)	22 854 (16.52%)	16 (1.23%)	6 940 (5.02%)

From these three threads it was shown that by setting an inclusion criterion of all ten key categories I extracted a little over 1% of the comments, ranging from 3-6% of the total words. The actual word count between the threads ranged from 5264 – 11 352 but in total, fewer than 25 000 words were taken forward for analysis. We might consider how this compares to an extraction process based on the top ten keywords: just two comments in the first discussion thread included all ten keywords and they were included in key comment extraction based on the top ten semantic categories. In the second thread, three comments included all ten keywords and in the third discussion thread, none of the comments incorporated all ten keywords. Again, the researcher must make a decision here as to what constitutes a suitable sample. Though the sample incorporating all ten key categories constitutes just ~1% of the overall comments we know that this particular 1% incorporated the key themes of the discussion and that the identification of those key themes is statistically valid. In the first discussion thread, relaxing the criteria to include any eight of the top ten key

semantic categories would extract a sample that numerically, represents 10% of the overall number of comments made. Determining the size of the sample will depend on the research question the researcher is looking to answer and the scope of data they are looking to include.

5 Discussion

Semantic annotation identifies key categories and offers a different perspective on language data to keywords in large datasets where prevalent themes are discussed using multiple terms. In a discourse that is shown to be creative and in which terminology is continually evolving (from 'global warming' to 'climate change' for example) relying on specific terms can be limiting. Researchers interested in the climate change debate would not be surprised by the nature of the key semantic categories revealed in this data. But there is an indication that such analysis can help us understand what is really at the heart of such discussions. As a consideration of the effect of online journalism and the changing roles of journalists and their readership, such data can be examined to determine the ways in which users shape the content and focus of articles that are produced online. Equally, we can explore how the subject of the article can invoke thematic shifts in the broader discourse around climate change. Secko *et al.* (2011) explored the relationship between the 'core audience narrative' and the 'core journalist narrative' but acknowledged the limitations of the type of analysis, "not possible for the data set as a whole" (p.819). Identifying key semantic categories provides a more cursory account of both the article and its discussion thread, enabling researchers to conduct cross-case comparisons.

This preliminary analysis of three user comment threads is not going to tell us anything about the nature of user comment threads in general, nor represent the discourse around climate change of those who comment on newspaper articles. As Scott (1997) has observed, such

analysis does not serve to “characterise a language or a genre, but a language event”. Online discussion threads are often dominated by a small number of contributors and are limited in how they represent public debates about climate change. In fact, in the first discussion thread for example, ten users posted 28% of the comments. By analysing the content and the reference to usernames, we find that equally those users with the highest numbers of comments were also the most often cited within the discussion: by reference to their username with the ‘@’ prefix, for example. The implications of this aspect for the deliberative potential of such discussion threads are considered in Collins and Nerlich (forthcoming).

6 Summary

The identification of key categories through semantic annotation incorporated more of the detail of the data than basic keyword analysis. The key categories identified in the data were not surprising but did highlight more features of the debate than standard keyword analysis, such as the prevalence of the discussions around scientific evidence and causality, as well as the claim/counter-claim nature of the discussion. I introduced a method of identifying key comments based on semantic annotation that can be used to extract a sample and allow the researcher to look more closely at key categories in context. I have explained how at each stage of this process the researcher must make certain decisions and set parameters in determining the nature and size of their sample, which can be adjusted to reflect their research aims. Semantic annotation is comparable to the manual thematic analysis methods more traditionally implemented in the discourse-based exploration of such data, demonstrating how computational processes can be used to make such methods more data-driven. I would suggest that researchers build on the advantages that have been shown in its application, but nevertheless consider alternative categorisation systems that could be applied

in a similarly computational manner in order to tackle the challenge of the breadth of online discourse data.

Acknowledgements

The author gratefully acknowledges funding from the Economic and Social Research Council, grant number: RES-062-23-1256.

References

Agirre, Eneko *et al*. (2009): "A study on similarity and relatedness using distributional and WordNet-based approaches." In Proceedings of NAACL-HLT 09, Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational linguistics. Boulder, USA: 19–27.

Archer, Dawn *et al*. (2003): "Developing an automated semantic analysis system for Early Modern English." In Archer, Dawn *et al*. (Eds.): *Proceedings of the Corpus Linguistics 2003 Conference*. UCREL technical paper number16. UCREL, Lancaster University: 22–31.

Baker, Paul *et al*. (2008): "A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press." *Discourse & Society* 19/3: 273–306. doi: 10.1177/0957926508088962.

Bassi, Erica (2010): "A contrastive analysis of keywords in newspaper articles on the 'Kyoto Protocol'". In Bondi, Marina/Scott, Mike (Eds.): *Keyness in Texts.* Studies in Corpus Linguistics 41. Amsterdam: John Benjamins Publishing: 207–218.

Bauer, Laurie/Nation, Paul (1993): "Word Families". *International Journal of Lexicography* 6/4: 253–279.

Berber Sardinha, Tony (1999): "Using keywords in text analysis: Practical aspects." *DIRECT Working Papers* 42. Available at: http://www2.lael.pucsp.br/direct/DirectPapers42.pdf. Accessed 21st June 2014.

Cheng, Winnie/Lam, Phoenix. W. Y. (2013): "Western perceptions of Hong Kong ten years on: a corpus-driven Critical discourse study." *Applied Linguistics* 34/2: 173–190. doi: 10.1093/applin/ams038.

Coe, Kevin/Kenski, Kate/Rains, Stephen A. (2014): "Online and Uncivil? Patterns and determinants of incivility in newspaper website comments." *Journal of Communication.* Available at: **http://onlinelibrary.wiley.com/doi/10.1111/jcom.12104/abstract**. doi: 10.1111/jcom.12104.

Collins, Luke C./Nerlich, Brigitte (forthcoming). "Do online user comments provide space for deliberative democracy? A corpus-based analysis of the climate change debate." *Environmental Communication.*

Culpepper, Jonathan (2009). "Keyness: words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet." *International Journal of Corpus Linguistics* 14/1: 29–59. doi: 10.1075/ijcl.14.1.03cul.

Fellbaum, Christiane (ed.) (1998): *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Firth, John (1968): "A synopsis of linguistic theory 1930-1955". In Palmer, Frank (ed.) *Selected Papers of J. R. Firth 1952-59.* Bloomington: Indiana University Press: 1–32.

Gabrielatos, Costas *et al*. (2012): "The peaks and troughs of corpus-based contextual analysis." *International Journal of corpus Linguistics* 17/2: 151–175. doi: 10.1075/ijcl.17.2.01gab.

Garside, Roger (1987): The CLAWS Word-tagging System. In Garside, Roger *et al*. (Eds.): *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Grundmann, Reiner/Krishnamurthy, Ramesh (2010): "The discourse of climate change: a corpus-based approach." *Critical Approaches to discourse Analysis across Disciplines* 4/2: 113–133.

Halliday, Mike A. K. (1994): *An Introduction to Functional Grammar.* Second edition. London: Edward Arnold.

Hermida, Alfred/Thurman, Neil (2008): "A clash of cultures: The integration of user-generated content within professional journalistic frameworks at British newspaper websites." *Journalism Practice* 2/3: 343–356. 10.1080/17512780802054538.

Hoey, Michael (1991): *Patterns of Lexis in Text.* Oxford: Oxford University Press.

Kaur, Sheena (2010): A corpus-driven contrastive study of girls' and boys' use of vocabulary in their free topic choice writing in Malaysia and the UK. Unpublished PhD thesis, Lancaster University.

Kehoe, Andrew/Gee, Matt (2012): "Reader comments as an aboutness indicator in online texts: introducing the Birmingham blog corpus." *Studies in Variation, Contacts and Change in English 12: Aspects of corpus linguistics: compilation, annotation, analysis.* http://www.helsinki.fi/varieng/series/volumes/12/kehoe_gee/. (Accessed 25th June 2014).

Koteyko, Nelya/Jaspal, Rusi/Nerlich, Brigitte (2013): "Climate change and 'climategate' in online reader comments: a mixed methods study." *The Geographical Journal* 179/1: 74–86. 10.1111/j.1475-4959.2012.00479.x

L'Hôte, Emilie/Lemmens, Maarten (2009): "Reframing treason: Metaphors of change and progress in New Labour discourse." *Cognitextes* 3. Available at: http://cognitextes.revues.org/248 Accessed 20th June 2014.

Lau, Jey Han/Baldwin, Timothy/Newman, David (2013): "On collocations and topic models." *ACM Transactions on Speech and Language Processing (TSLP) – Special issue on multiword expressions: From theory to practice.* 10/3: article 10 1-14. 10.1145/2483969.2483972.

Leech, Geoffrey (1997): "Grammatical tagging." In Garside, Roger/Leech, Geoffrey/McEnery, Tony (Eds.): *Corpus Annotation: Linguistic Information from Computer Text Corpora.* London: Longman: 19–33.

Manosevitch, Edith/Walker, Dana (2009): "Reader comments to online opinion journalism: A space of public deliberation." Paper presented at the *10th International Symposium on Online Journalism, 17-18th April, 2009.* Austin, TX.

Milioni, Dimitra/Vadratsikas, Konstantinos/Papa, Venia (2012): "'Their two cents worth': A content analysis of online readers' comments in mainstream news outlets." *Observatorio (OBS*)* 6/3: 21–47.

Murphy, Shane (2006): "Now I am alone: A corpus stylistic approach to Shakespearian soliloquies." In Gabrielatos, Costas/Slessor, Richard/Unger, Johann W. (Eds.): *Papers from the Lancaster University Postgraduate Conference in Linguistics and Language Teaching* Vol. 1. Lancaster University, 15 July 2006: 66–85. Available at: http://www.ling.lancs.ac.uk/pgconference/v01/Volume01.pdf. Accessed 20th June 2014.

Nerlich, Brigitte/Forsyth, Richard/Clarke, David (2012): "Climate in the news: How differences in media discourse between the US and UK reflect national priorities." *Environmental Communication: A Journal of Nature and Culture* 6/1: 44–63. 10.1080/17524032.2011.644633.

Partington, Alan (2004): "'Utterly content in each other's company': Semantic prosody and semantic preference." *International Journal of Corpus Linguistics* 9/1: 131–156. doi: http://dx.doi.org/10.1075/ijcl.9.1.07par.

Piao, Scott S. L., *et al*. (2004). "Evaluating lexical resources for a semantic tagger." In: *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004).* Volume II. May, 2004. Lisbon, Portugal: 409–502.

Potts, Amanda/Baker, Paul (2012): "Does semantic tagging identify cultural change in British and American English?". *International Journal of Corpus Linguistics* 17/3: 295–324. doi: 10.1075/ijcl.17.3.01pot.

Rayson, Paul (2003): Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. PhD thesis: Lancaster University.

Rayson, Paul (2008): "From key words to key semantic domains." *International Journal of Corpus Linguistics* 13/4: 519–549. doi: 10.1075/ijcl.13.4.06ray.

Rayson, Paul *et al*. (2000): "The REVERE Project: Experiments with the application of probabilistic NLP to Systems Engineering." In: *Proceedings of 5th International Conference on Applications of Natural Language to Information Systems*. 28-30th June, 2000. Versailles, France.

Rayson, Paul *et al*. (2004): "The UCREL semantic analysis system." Proceedings of the Workshop "Beyond Named Entity Recognition. Semantic Labelling for NLP Tasks" in association with the 4th International Conference on Language Resources and Evaluation (LREC). 26-28th May 2004. Lisbon, Portugal: 7–12.

Scott, Mike (1997): "PC analysis of key words – and key key words." *System* 25/2: 233–245.

Scott, Mike (1999): *WordSmith Tools Help Manual, Version 3.0*. Oxford: Oxford University Press.

Scott, Mike (2007): *WordSmith Tools Help Manual, Version 4.0.* Available at: http://www.lexically.net/downloads/version4/wordsmith.pdf. Accessed 10th June 2014.

Scott, Mike/Tribble, Chris (2006): *Textual Patterns.* Amsterdam: John Benjamins.

Secko, David .M. *et al*. (2011): "The Unfinished Science Story: Journalist-audience interactions from the Globe and Mail's online health and science section." *Journalism* 12/7: 814–831. doi: 10.1177/1464884911412704.

Suler, John (2004): "The Online Disinhibition Effect." *Cyberpsychology and Behavior* 7/3: 321–326.

Tausczik, Yla R./Pennebaker, James W. (2010): "The psychological meaning of words: LIWC and computerized text analysis methods." *Journal of Language and Social Psychology* 29/1: 24–54.

Tognini-Bonelli, Elena (2001): *Corpus linguistics at work.* Amsterdam: John Benjamins.

Williams, Raymond (1983): *Keywords: A Vocabulary of Culture and Society*. Second Edition. London: Fontana.

Wilson, Andrew (1993): "Towards an integration of content analysis and discourse analysis: the automatic linkage of key relations in text." UCREL Technical Paper 3 Linguistics Department: Lancaster University.