Topically-focused Blog Corpora for Multiple Languages

Andrew Salway[1], **Dag Elgesem**[2], **Knut Hofland**[1], **Øystein Reigem**[1], **Lubos Steskal**[2]
[1]Uni Research, Bergen, Norway [2] University of Bergen, Norway
{andrew.salway, knut.hofland, oystein.reigem}@uni.no
{dag.elgesem, lubos.steskal}@uib.no

Abstract

This paper describes the construction of three corpora, intended for use in social science research, comprising English-language, French-language and Norwegian-language blogs related to the topic of climate change. The approach, techniques and lessons learnt should be applicable for creating other topically-focused blog corpora.

1 Introduction

Since the 1990s blogs have emerged as an important medium in which users can easily create and share content on the Internet. The emergence of the blogosphere has brought changes to the online public sphere, to the role of the mainstream media, to the production, contestation and dissemination of scientific knowledge, and to political deliberation. As a site for large-scale discourses about socially-relevant issues, the blogosphere has received considerable attention from social scientists during the last decade (Rettberg, 2013; Bruns and Jacobs, 2006).

Important questions relate to the democratic potential of blogs, i.e. whether they do indeed provide a new platform for open democratic participation (Benkler, 2007), or whether a minority of blogs get most of the attention (Hindman, 2008). Researchers have studied the roles of blogs in connection with political campaigns (Adamic and Glance, 2005; Bruns and Jacobs, 2006; Moe, 2011) and controversial political issues, like climate change, where the diffusion of information may influence the formation of opinions (Sharman, 2014; Elgesem et al., 2015). One aspect is whether the linking practices of bloggers contribute to the polarization of online political debate and the fragmentation of the online public sphere (Sunstein, 2008). Also, the relationship between mainstream media and blogs has been studied, e.g. to see whether blogs influence the audience's attention to news (Bruns, 2005; Leccese, 2009; Elgesem et al., 2016).

Despite the great interest in the content of the blogosphere, there is a lack of commonly available large-scale blog corpora to support empirical research. Most blog corpora created for social science research have been relatively small since they were concentrated on what were perceived to be the most important blogs for certain research questions (e.g. Adamic and Glance, 2005; Song et al., 2007; Sharman, 2014). Larger blog corpora have been created but these were not focused on particular topics or were not designed to support social science research (e.g. Glance et al., 2004; Bansal and Koudas, 2007; Kehoe and Gee, 2012; Meinel et al., 2015).

One exception is our previous development of a large climate change blog corpus (Salway et al., 2013; Elgesem et al., 2015). However, the method used in that work was somewhat ad hoc in its selection of blogs when crawling. The method's reliance on human judgment means that it is hard to replicate, i.e. in order to update the corpus, create corpora for other languages and topics, and critique it as part of social science methodology. Further, crawling-based methods may be problematic when a topically-defined area of the blogosphere is fragmented. It may be expected that there are few, if any, connections between some communities in the climate change blogosphere.

This paper describes the construction of large topically-focused blog corpora which are intended for use in social science research. The topic is climate change and the languages are English, French and Norwegian. It is hoped that the approach and techniques can be usefully replicated for other topics and languages. By topically-focused we mean a corpus that contains "all" blogs related to a socially-relevant issue, like climate change. By providing, as far as possible, an unbiased and comprehensive collection of relevant blogs, including core blogs and the

broader discourse around them, such a corpus supports a variety of social science research.

2 Task definition and approach

An important aspect of the blogosphere is the interaction between bloggers as evidenced by their linking patterns. Thus blog corpora should contain data about hyperlinks as well as the main text component of every blog post. The date of each blog post is needed for investigating the development of blog communities and information diffusion. A blog contains various pages, e.g. a homepage, archive pages, and posts. We seek to harvest all the posts for each chosen blog, but not other pages. So, in simple terms, the task at hand is to create a corpus containing all posts – with text, link and date data – from all blogs in a chosen language and that relate to a chosen topic. In this section we describe and discuss how the notions of blog, topic and language are defined and operationalized in our approach.

Blogs are commonly understood to be discussion or informational websites with posts presented in reverse chronological order. For practical reasons we define a blog to be a website that is produced using one of several blog authoring platforms, and, more specifically, a website that mentions the platform in its domain name. The platforms were chosen based on search engine results for queries in order to establish which platforms dominated searches for terms related to the topic. Searching for "climate change" and "global warming" in blogs showed that around 95% of all hits came from blogs on the WordPress, Blogspot and TypePad platforms. No other platform had more than 1% of the hits. The same platforms dominated for Norwegian; for French, OverBlog swapped with TypePad.

This operational definition of blog – as a website that mentions a blog platform in its domain name – means it is trivial to identify blogs in a consistent way when searching and crawling. Selecting only a few blog platforms means that we can afford to optimize data extraction techniques in a platform-specific way. The obvious negatives are that we miss blogs on other platforms, and also blogs that are produced on the chosen platforms but that only use a domain name that does not contain the platform name. Results from the work reported in this paper suggest that some of the most important blogs are known by such domain names (see 4.4).

A topic like climate change is very broad and rather nebulous. People may blog about climate change from scientific, political and social perspectives, within which there are competing viewpoints. Blogs further vary in how they focus on the existence of climate change, its causes, its effects and ways in which to mitigate or adapt. The discussions may be in the context of local geographic areas, countries or the whole world. Some blogs will be specifically about climate change issues, but many other scientific, political and socially-concerned blogs mention it.

For blog corpora to be used in social science research it is important to minimize bias towards particular people, perspectives and viewpoints. As far as possible the method for selecting blogs should be transparent. Thus, we chose to define the topic of climate change with only a few generic terms, i.e. for English, "climate change", "global warming" and "greenhouse effect". These terms were chosen following the work of Schmidt et al. (2013) who conducted an extensive review of research into climate change communication: they considered the three terms to refer to the same phenomena and used them, and variant forms like 'climat* NEAR chang*', to select relevant newspaper stories. Whilst query expansion methods could be used to add many other search terms related to the topic, e.g. "sea level", "climate sceptic", "carbon tax", etc., we feel that this could introduce unaccountable bias into the selection of the material.

In our approach the search terms are used with search engine APIs in order to identify relevant blogs. In brief, the method retrieves blog posts containing the search terms and then selects blogs that have >1 posts containing >1 instances of a search term (see 3.1). This criterion is deliberately inclusive, i.e. it is intended to include blogs with only very few mentions of search terms because: (i) some blogs may focus on a specific aspect of the climate change debate without mentioning the generic terms very often; (ii) some blogs may be tangentially related to the climate change debate whilst still being of interest to some researchers. Researchers can later apply stricter criteria to select sub-corpora as necessary for specific research questions.

It is problematic to define and operationalize the concepts of nationality and national language varieties in the blogosphere. A blogger may write in their native language but be living and writing in the context of another country, or write in a lingua franca for an international audience. Some blogs have multiple contributors of different nationalities using different languages. Even if it was desirable to classify blogs according to nat-

ionality, practically it is not possible to reliably connect a blog to a country from its url, nor ascertain the nationality of a blogger.

Our approach is to create English-language, French-language and Norwegian-language corpora, without associating blogs with countries or language varieties. So, for example, an English-language corpus may include blogs written in US, Australian and British varieties, etc., and bloggers of any nationality, including some writing in English as a second language. Language identification is achieved with language codes when querying search engines (3.1), and subsequently an off-the-shelf tool (3.3.3).

3 Pipeline

3.1 Identification of relevant blogs

For each language a set of potentially relevant blog posts was gathered from repeated querying of three search engine APIs (Google, Bing and Yahoo). It seemed appropriate to use multiple APIs to reduce bias from any single one, although because Bing and Yahoo allowed more results to be returned it may be that they have a bigger influence than Google. That said, it must be noted that the search engine APIs are "black boxes" to us, i.e. we cannot know how they determine result sets and there is a risk of "filter bubble" effects (Pariser, 2011).

Queries specified a search term, a blog platform and a language code. For English the terms were "climate change", "global warming" and "greenhouse effect"; these were translated into French (three terms with five inflections) and Norwegian (four terms with 12 inflections). It could be that the search engine APIs would expand search terms into their inflected forms, but it seems safer to be explicit, and perhaps it helps to reach further down the list of potential results.

Querying was done daily for 12 weeks from early June 2014 and the rate of previously unseen posts and blogs in the results was monitored. New posts in the results were due both to bloggers writing new posts, and to search engines re-ranking older ones. Since search engines limit the number of results returned (100-1000 per query), after two weeks the set of query terms was expanded with n-grams containing the initial search terms and a function word, e.g. "of climate change". For each initial search term we took the 10 most frequent n-grams from the posts returned from the search engines at that point. This allowed us to reach much further down the search engines' results lists. Whilst this helped to retrieve many more relevant posts, it also meant that for Norwegian, and to some extent French, some retrieved posts were rather tangential to the topic. It was also noted that about 20% of the posts returned for Norwegian were actually in Danish and had to be removed: this was done using a list of frequent Danish words that are rare in Norwegian. The total cost for using the APIs was approximately $2000. Table 1 shows for each language: the search terms, blog platforms and the number of blog posts that were retrieved.

English (WordPress, BlogSpot, TypePad) climate change, global warming, greenhouse effect → 95,662 posts
French (WordPress, BlogSpot, OverBlog) changement climatique, changements climatiques, réchauffement climatique, effet de serre, effets de serre → 68,853 posts
Norwegian (WordPress, BlogSpot, TypePad) drivhuseffekt, drivhuseffekten, global oppvarming, globale oppvarmingen, klimaendring, klimaendringen, klimaendringene, klimaendringer, klimaforandring, klimaforandringen, klimaforandringene, klimaforandringer → 8,973 posts (after Danish removed)

Table 1: The search terms and blog platforms for each language, and the number of posts returned from querying search engine APIs.

The sets of retrieved blog posts were used to determine which blogs should be harvested. Data was generated about the occurrence of the search terms in the retrieved posts, and hence in the blogs which they came from. The main text of each post was extracted with *jusText* (Pomikálek, 2011) and concordance lines of the search terms were inspected in order to identify and remove striking examples of duplicates due to boilerplate and spam posts. This gave us "good enough" text extraction for this stage of the process.

Table 2 shows how many blogs had >0, >1, >2 and >3 posts that contained >0, >1, and >2 instances of search terms. For example, for English there were 5563 different blogs for which we had >1 posts that contained >1 instances of search terms. Drawing on domain expertise, the sets of blogs relating to the different frequencies were inspected in order to decide appropriate thresholds. In order to favor broad inclusion it was decided to harvest all posts from blogs for which we had gathered >1 posts containing >1 instances of key terms, i.e. 5563 English, 2088 French and 128 Norwegian blogs.

The values for "total posts" in Table 2 are lower than in Table 1 because text was not extracted from all posts; either there was no text in

the post, or text extraction failed. After text extraction not all posts contained a search term, e.g. only 67,979 out of 84,536 English posts did. From preliminary inspection it seems that this is because search terms only occurred in the boilerplate of some blog posts, and hence not in the extracted texts. It could also be a sign of query expansion by the search engines.

Search terms	Total posts	Blogs >0 post	Blogs >1 post	Blogs >2 posts	Blogs >3 posts
	English				
Total	84536	27873	7205	3995	2762
>0	67979	25190	6515	3541	2391
>1	56806	21231	**5563**	2998	2042
>2	46584	18007	4633	2493	1730
	French				
Total	52029	13838	4552	2716	1931
>0	35578	12732	3926	2217	1526
>1	17655	6470	**2088**	1213	845
>2	10839	4187	1318	754	512
	Norwegian				
Total	7194	613	505	293	224
>0	2794	1393	337	172	119
>1	943	470	**128**	65	42
>2	477	268	67	26	18

Table 2: The data used to select blogs, i.e. the occurrence of search terms in posts, and the occurrence of these posts in different blogs.

3.2 Harvesting, pre-processing and filtering

The aim was to harvest all posts from the selected blogs that were posted up until the end of 2014. The harvesting script was customized for each blog platform but in general it started at each blog's homepage, followed links to archive pages and got the urls from all links in them. Each url was tested to make sure it had the recognized features for a blog post and that it was from 2014 or earlier (the year is given in the url). The harvesting script was improved iteratively, and rerun, as we learnt more about the idiosyncrasies of each blog platform. For each post the html was run through *ftfy* (Speer, 2016) to address encoding issues, and the time of harvesting was recorded.

The same blog post can be referred to with different urls which causes duplication in the corpus. It also causes problems for analyzing the network between blog posts, i.e. when hyperlinks point to the same post using different urls. Normalization of urls used manually created look-up tables to resolve alternative domain names. Rules were applied to standardize character encoding, the use of www and http, and platform-specific formatting variants. For blog posts with the same urls after normalization, we used the first html file and kept a record of the urls that had been normalized to it. In the English and French material 4.8% of posts had duplicates (nearly always just one); 3.3% for Norwegian.

3.3 Data extraction from html

3.3.1 Boilerplate removal (aka text extraction)

To support social science research it is important to extract the main text of each blog post as accurately as possible, i.e. so that it is then possible to analyze and compare what was written where and when. This contrasts with some web as corpus initiatives in which the corpus may be treated as a bag of sentences in order to remove duplicates (Biemann et al., 2013).

We evaluated two general text extraction solutions – *jusText* (Pomikálek, 2011) and *Alchemy* (www.alchemyapi.com) – on a sample of 1000 posts. For about 20% of posts there was either text missing or extraneous text included. Thus it was considered necessary to develop our own text extraction tool which could take advantage of the fact that the posts it processed came from specific blog authoring platforms.

We assume that the main text of the blog post is continuous and that each blog platform has a certain amount of regularity in how html sequences indicate the start and end of the main text; cf. the BTE algorithm and jusText algorithm reviewed and combined by Endrédy and Novák (2013). For each platform, a set of heuristics – based on html cues – was iteratively developed to identify the start and end points of the main text within an html file. This involved counting frequent elements, manual inspection of html files, and trial and error application of heuristics in which the matching heuristics were recorded and counted, so that the most useful ones became apparent.

The main text is taken to be all lines of html from the first instance of a start cue until the first instance of an end cue. From the selected lines all html tags, and other html sequences, were stripped except link, paragraph and break markers, ensuring white space was maintained. Finally we removed multiple whitespace, converted html entities to characters, and substituted a uniform marker for paragraphs and breaks.

When run over all harvested posts in the three corpora the start and end cues succeeded in matching for 99.7% of all posts, i.e. something

was extracted as main text for nearly all of the posts. The quality of the extraction was evaluated with a set of 1463 randomly selected English-language posts, all from different blogs. The evaluator determined that there were only 11 posts (< 1%) in which part of the main text was missing. There were 72 (5%) posts with inappropriate text included at the beginning, 1 in the middle and 48 (3%) at the end.

3.3.2 Further boilerplate mark-up by 5-grams

In the case of a blog corpus, unwanted boilerplate text can manifest as near-duplicate paragraphs in the extracted article text for many posts within a blog, e.g. a slogan for the blog, or a request for donations. However, it should not be assumed that this kind of boilerplate will appear consistently on all posts in a blog since it may change during the life of a blog, or between different batches of a harvest.

Even if we are confident of getting a high precision rate in identifying blog-specific boilerplate within the previously extracted text, it seems better that we mark it up, rather than delete it and risk destroying some relevant material. This means that researchers can decide later what to include for their investigations; for some researchers blog-specific boilerplate might even be an object of study.

Our approach to marking-up blog-specific boilerplate text is based on identifying a set of suspicious 5-grams for each blog, i.e. 5-grams that occur on more than a certain percentage of posts. Through analysis and trial and error we determined a threshold of 15%. Because we had harvested posts in different batches it was important to keep the threshold percentage quite low, i.e. $\geq$ 15% (and frequency $\geq$ 10), in case boilerplate text changed from batch to batch. However we noticed that some genuine 5-grams did occur on more than 15% of a blog's posts, due to the idiosyncratic writing style of some bloggers. We also note that some boilerplate paragraphs may consist of fewer than 5 words but we tolerate these because 4-grams, and less, are too common in normal text.

Thus any paragraph for which 50% or more of the words comprise suspicious 5-grams was marked as boilerplate. We say 50% to capture boilerplate lines in which some content may vary, like a date or a name: it seems unlikely that 50% of a real paragraph would be made up of common 5-grams. Manual evaluation (see 3.3.1) showed that in 23 (9%) of 258 posts with boilerplate marked-up, one or more paragraphs had been incorrectly marked-up as boilerplate. Whilst 9% is quite high, it is likely that only a small part of each post was incorrectly marked as boilerplate. Further, boilerplate was only marked-up at all on about 20% of all blogs.

3.3.3. Posts in wrong language

For several reasons it is possible that a blog in a corpus for one language contains some posts that are in a different language. Firstly, as noted previously, blogs may contain posts from different contributors or be written by someone in more than one language. Another problem, especially relating to English, is that our search terms may occur on blogs that are all in another language, e.g. as part of a quote or a scientific citation. Due to our low threshold for selecting blogs (see 3.1) a non-English blog would be selected for the English corpus if it had two posts each with two mentions of any English search terms. Further issues may arise from the reliance on search engines' language codes.

We anticipate that different users of the corpora will have different ideas about what should count as an English/French/Norwegian-language blog; and for some researchers the multilingualism of blogs could be of interest. Thus it does not seem appropriate to delete material but rather to record measures of how many posts in each corpus, and in each blog, are in the target language.

Every blog post was run through *langid.py* (Lui and Baldwin, 2012) and a Boolean value was recorded according to whether the post was most likely to be in the target language of its corpus or not. The results of correct language posts by corpus were: English, 96%; French, 81%; Norwegian, 89%. It may be that some blogs are bilingual, e.g. Canadian blogs in French and English, but this remains to be explored.

For each blog the percentage of posts in the correct language was calculated. For analysis, a threshold of 85% posts was taken to mean that a blog was principally in the correct language; the threshold was based on manual inspection of blogs with high numbers of incorrect language posts. This gave the following estimate of blogs in the correct language: English, 96%; French, 86%; Norwegian, 87%.

3.3.4 Link extraction

For each post we stored a set of article links, i.e. the urls (normalized as per 3.2) pointed to from all links found in the main text (article). These links are also marked up in the main text because

when analyzing linking patterns in social media it is important to consider the text around links.

Data was also stored to facilitate network analysis based on blogroll links. A blogroll typically appears in a sidebar on all posts of a blog and includes links to other blogs that the blogger is assumed to have some affinity with. We explored extracting these links based on html structure but did not see sufficient regularity. Instead, going blog-by-blog, for each link we store the percentage of the blog's posts that it appears on (instances of links appearing within article text are not counted). A high percentage value for a link that points to another blog may be assumed to indicate a blog roll link (see 4.3).

3.3.5 Date extraction

On the four chosen blog platforms the url for a post normally includes values for month and year; WordPress urls also contain a value for day. For now we simply record these values for a blog post's date. For 490 French OverBlog blogs date information was not available from urls. Work has been done to iteratively develop heuristics to extract date data, including day values, from html for all platforms; see the approach to text extraction, 3.3.1. Preliminary evaluation is encouraging based on comparing month and year values extracted from the html with those from urls. However, date data from html has not been incorporated into the corpora yet.

4 Analysis of the corpora

Table 3 records the content of the three corpora. It counts posts considered to be in the target language of each corpus (see 3.3.3), and words in the main texts of posts, excluding paragraphs marked-up as boilerplate (see 3.3.1 and 3.3.2). The total number of blogs is 123 (1.6%) less than stated in Table 2: this is due to a combination of harvesting and text extraction problems, including the fact that some blogs were no longer available for harvest.

	Blogs	Posts	Words
English	5497	10,539,575	4,837,481,377
French	2033	2,335,174	1,224,657,286
Norwegian	126	46,775	21,212,686

Table 3: The content of the three climate change blog corpora.

Figure 1 shows, for the English corpus, the cumulative total percentage of posts in the corpus against the rank of each blog by number of posts. For example, it shows that the top 20 blogs account for 10% of the corpus by posts, and the top 200 blogs for 40%. The weighting towards top ranked blogs in the French and Norwegian corpora is even greater. From initial inspection it seems that most of these very large blogs are only loosely related to climate change, if at all. Users of the corpora may consider excluding some of them from their investigations.

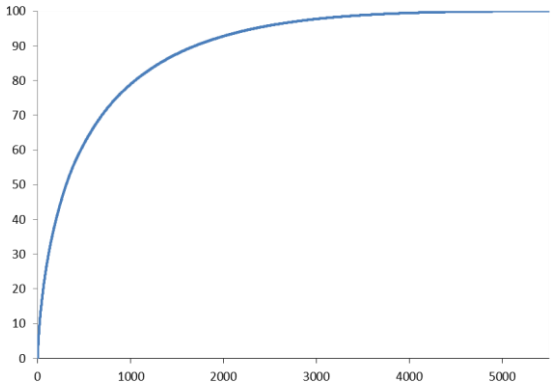

Figure 1: Cumulative total percentage of posts in the English corpus against rank of blog by number of posts.

4.1 Text analysis

As a first step to understand the extent to which the English corpus relates to the topic of climate change, Table 4 gives a view of the distribution of the three search terms used to select blogs. For each term it gives the percentage of blogs that have at least one post containing it, and the percentage of all posts containing it. The fact that 99.2% of blogs, and not 100%, contain any search term suggests minor problems with either harvesting or with text extraction.

The table also gives '% pwc' which sums the word count for all posts containing the term and shows this as a percentage of all words in the corpus. Thus about 14% of the corpus (by word count) comprises blog posts with at least one mention of a search term. Work is ongoing to determine how many further posts contain other terms related to climate change.

	freq.	% blogs	% posts	% pwc
Any term	2,415,596	99.2	6.4	13.9
climate change	1,486,549	96.5	4.8	11.6
global warming	900,918	96.1	3.3	8.7
g'house effect	28,129	47.6	0.1	0.6

Table 4: The distribution of search terms in the English-language corpus.

Another view of the distribution of the terms was obtained by considering, for each blog, the percentage of posts that contain at least one term. This showed that a large number of the blogs appear to be only tangentially related to the topic, although an expanded set of terms needs to be considered before conclusions are made. Some 1041 blogs out of 5497 have only 0-2% of posts containing a search term. A further 1907 blogs have 2-10% posts with search terms, and 742 blogs have 10-20%. The remaining 1807 blogs are evenly spread between 30-100%.

4.2 Distribution of dates

For a temporal view, Figure 2 shows the rate of posting increasing steadily year-by-year in each corpus. Of course it is possible that there is a recency effect since search engines were used to identify relevant blogs: perhaps some blogs that were only active several years ago were missed.

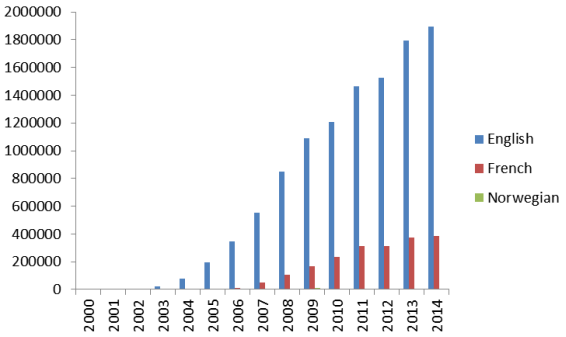

Figure 2: Number of blog posts per year.

By considering the date of the earliest post in each blog, Figure 3 shows a peak for blogs being started in 2009, with a fairly steady decline since then. When the date of the most recent post in each blog was examined it appeared that 56% of English blogs were still active at some point in 2014; French, 61%; Norwegian, 57%.

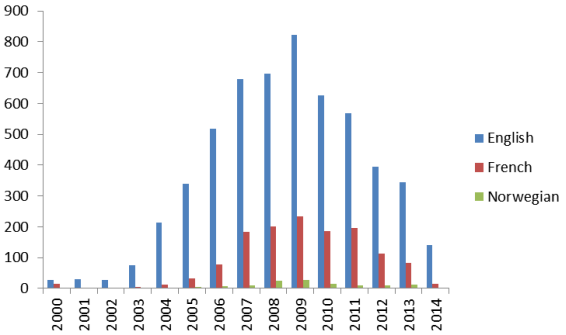

Figure 3: Number of blogs starting per year, i.e. the year of their earliest post.

4.3 Network analysis

As described in 3.3.4, for each non-article link found on a blog we calculated the percentage of posts in that blog that it appeared on. The assumption is that links pointing to the same url from most posts on a blog are blogroll links. Considering the distribution of these percentage values for the English-language corpus we see that non-article links tend to occur either on very few of a blog's posts, or on most of them: 70% of the 855,778 links occur on < 10% of the posts in the blogs they occur on; 24% occur on > 90% of the posts in the blogs they occur on. This leads us to take non-article links occurring on >90% of a blog's posts as blogroll links.

To analyze the network structure of the corpus this set of blogroll links was filtered to keep only those pointing to a blog that we had harvested. Then a directed blog network was created where there is a directed edge from blog A to blog B if there is a blogroll link from A to B. As is common in web networks, the degree distribution is power law like (Broder et al., 2000), ranging from 0 to 68. A few blogs have most of the inlinks (links pointing to the blog), and most blogs have very few or none.

To visualize the network all nodes with an indegree < 5 were removed. Figure 4 shows a network visualization made in Gephi (Bastian et al., 2009) with the ForceAtlas 2 layout algorithm (Jacomy et al., 2014): this tends to cluster highly interconnected nodes and repel weakly connected ones. On top of this, a modularity based community detection algorithm (Blondel et al., 2008) clustered the network into four densely connected groups distinguished by the shade of grey. Note, the size of a node reflects its indegree. Manual inspection of core blogs in each of the four clusters suggested that one cluster comprised mostly skeptical blogs, one acceptor blogs, and one blogs concerned primarily with economic issues; the fourth cluster was less coherent.

4.4 How much of the climate change blogosphere was captured?

Generally it may be assumed that a blog's indegree reflects its importance. Hence, one way to assess the coverage of the corpora, and to fill gaps, is to use blogroll links as a source of information about what are important blogs. Here we take some preliminary steps in this direction.

For the blogroll links, filtered by platform, from the English-language corpus (see 4.3) the

in-degree for each linked-to blog was counted, including blogs not in the corpus. Table 5 shows the number of blogs with different in-degrees, and the percentage of these blogs that are present in the corpus. Assuming a minimum in-degree of 25 to mean that a blog is important, our method retrieved 22 (88%) of the important climate change blogs with the chosen platforms in their domain names. The missing blogs are about politics in general, rather than climate change specifically (dissectleft, nicedeb, gatewaypundit).

Minimum in-degree	Blogs in network	Blogs in corpus	"recall" %
1	19010	1255	6.6
2	3279	638	19.5
5	560	234	41.8
10	159	95	59.7
15	76	51	67.1
20	40	32	80.0
25	25	22	88.0

Table 5: Assessing coverage of the English-language corpus, where a high in-degree is assumed to indicate an important blog.

By taking blogroll links from the whole corpus the analysis is swamped by links from blogs that are mostly peripheral to the climate change blogosphere. To address this, we also examined all the blogroll links from six well known climate change blogs reflecting different perspectives and viewpoints (joannenova, realclimate, wattsupwiththat, tamino, climatechangeaction, climate-connections). In this case links to all blogs judged to be about climate change were included, i.e. not only blogs explicitly on the chosen platforms. This gave 71 blogs that are not present in the English-language corpus. Seven of these have 'wordpress' or 'blogspot' in their domain name, so could and should have been captured by our selection method.

Of the 64 missing blogs without any platform mentioned in their domain name, 17 had an in-degree (from the corpus) >25 and hence may be considered crucial omissions, although they vary in the extent to which they are about climate change or more general environmental and political topics. It is interesting that 17 out of 64 missing blogs not explicitly on any platform should be so important. This compares with 22 out of 5497 blogs explicitly on a chosen platform with an in-degree >25 (Table 5). This suggests a strong tendency for important blogs to use a domain name that does not include any blog platform. Hence it seems that, at least for some social science investigations, our blog selection method would have to be extended in order to capture more of the important blogs. This could be done by systematically using in-degree data to crawl from the initial corpus. However, human judgment would be required to determine what linked-to websites were blogs, and another test would be required for topic relevance.

Figure 4: The directed blog network, based on blogroll links, from the English-language corpus. This shows most blogs are in one of four communities. The size of each node (blog) reflects its in-degree.

5 Closing remarks

It is not possible to make strong claims about how successful any method is in retrieving all blogs related to a topic. This is due to the fuzzy boundaries of topics and to the lack of a common definition of what constitutes a blog being related to a topic, rather than mentioning it in passing. However, the preliminary analyses in Section 4 allow us to say something about the efficacy of our approach and techniques. In brief, large-scale blog corpora were created with a reasonable amount of topical content, and intuitively correct temporal distribution and network structure.

To recap, the method relies on two main assumptions. Firstly, a blog is considered to be relevant if search engine APIs return >1 posts with >1 instances of the search terms which are chosen to be generic for the topic. This criterion could be considered to be too permissive, i.e. it includes blogs that are not really relevant to the topic. However, researchers have the option to apply stricter criteria and create sub-corpora that are better suited for specific research questions. The use of only a few generic search terms might mean that some niche sites that focus on a particular aspect of the topic are missed if they do not use the generic terms. We feel this is unlikely but it perhaps should be tested in future work.

Secondly, a website is considered to be a blog if it includes one of the selected blog platforms in its domain name. As discussed in 4.4 this leads to some important blogs being missed but this could remedied by selecting further blogs based on link data from the initial corpus and human judgement. Of course, a different approach would be to gather all blogs by crawling from an initial set of seed blogs, as we did in previous work (Salway et al., 2013; Elgesem et al., 2015). Two reasons seem to count against such an approach: (i) in fragmented blogospheres crawling may miss communities that are weakly connected to the rest; (ii) ensuring only topical blogs are included could entail downloading and analyzing a prohibitively large amount of most websites visited. Ideally, future work would make a systematic comparison of results from the two approaches and look to combine them.

Regarding the replicability and transparency of the method, a potential drawback is the reliance on search engine APIs whose ranking algorithms and query expansion techniques are unknown to us. Bias is mitigated to some extent by using multiple search engines, and by having a low threshold for what blogs are included (3.1). However, the ever changing nature of the algorithms counts against precise replicability.

We are currently preparing and documenting the corpora so that they can be made available for research purposes. In the first instance they will be released as they are currently, i.e. without removing 'further boilerplate' (3.3.2) and 'wrong language' material (3.3.3), and without adding further blogs (4.4). However, the corpora will include all the information needed to allow researchers to make their own decisions about how to customize the corpora to address specific research questions. Work is ongoing to integrate date data from html, and to extract data about comments. Also ongoing is work to further investigate the content of the corpora as part of social science investigations, e.g. to identify acceptor and sceptic communities and analyze the interactions between them, and to compare this between the different corpora.

Acknowledgements

This work was funded by the Research Council of Norway (VERDIKT program) and the Center for Big Data Analysis, Uni Research. We are very grateful for the technical support provided by Patcharee Thongtra and Eirik Thorsnes. Daniel Rognes and Samia Touileb helped with evaluation. Kjersti Fløttum and Anje Muller Gjesdal provided input regarding the content of the French corpus. Finally, many thanks to the three anonymous reviewers for their constructive feedback.

References

Lada A. Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 US election: divided they blog. *LinkKDD'05*: *Proceedings of the 3rd International Workshop on Link Discovery*:36-43.

Nilesh Bansal and Nick Koudas. 2007. BlogScope: A System for Online Analysis of High Volume Text Streams. *Procs. VLDB '07*:1410-1413.

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: an open source software for exploring and manipulating networks. Procs. Third International AAAI Conference on Weblogs and Social Media: 361-362.

Yochai Benkler. 2007. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press.

Chris Biemann, Felix Bildhauer, Stefan Evert, Dirk

Goldhahn, Uwe Quasthoff, Roland Schäfer, Johannes Simon, Leonard Swiezinski, and Torsten Zesch. 2013. Scalable Construction of High-Quality Web Corpora. *Journal for Language Technology and Computational Linguistics* 28(2):23-59.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 10, P10008.

Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. 2000. Graph structure in the web. *Computer Networks* 33:1, 309-320.

Axel Bruns. 2005. *Gatewatching: Collaborative Online News Production*. Peter Lang, New York.

Axel Bruns and Joanne Jacobs, eds. 2006. *Uses of Blogs*. Peter Lang, New York.

Dag Elgesem, Lubos Steskal, and Nick Diakopoulos. 2015. Structure and Content of the Discourse on Climate Change in the Blogosphere: The Big Picture. *Environmental Communication* 9(2):169-188.

Dag Elgesem, Ingo Feinerer, and Lubos Steskal. 2016. Bloggers' Responses to the Snowden Affair: Combining Automated and Manual Methods in the Analysis of News Blogging. *Computer Supported Cooperative Work (CSCW)* 25(2):167-191.

István Endrédy and Attila Novák. 2013. More Effective Boilerplate Removal – the GoldMiner Algorithm. *Polibits* (48):79-83.

Natalie S. Glance, Matthew Hurst, and Takashi Tomokiyo. 2004. BlogPulse: Automated Trend Discovery for Weblogs. *WWW 2004 Workshop on the Weblogging Ecosystem*.

Matthew Hindman. 2008. *The Myth of Digital Democracy*. Princeton University Press.

Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one* 9.6: e98679.

Andrew Kehoe and Matt Gee. 2012. Reader comments as an aboutness indicator in online texts: introducing the Birmingham Blog Corpus. *Studies in Variation, Contacts and Change in English* 12.

Mark Leccese. 2009. Online Information Sources of Political Blogs. *Journalism and Mass Communication Quarterly* 86(3):578-593.

Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. *Procs. ACL 2012*: 25-30.

Christoph Meinel, Justus Broß, Philipp Berger, and Patrick Hennig. 2015. *Blogosphere and its Exploration*. Springer-Verlag, Berlin.

Hallvard Moe. 2011. Mapping the Norwegian Blogosphere: Methodological Challenges in Internationalizing Internet Research. *Social Science Computer Review* 29(3):313-326.

Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Jan Pomikálek. 2011. Removing boilerplate and duplicate content from web corpora. PhD thesis, Masaryk university, Faculty of Informatics, Brno, Czech Republic. Software: https://pypi.python.org/pypi/jusText

Jill Walker Rettberg. 2013. *Blogging.* Polity Press.

Andrew Salway, Samia Touileb, and Knut Hofland. 2013. Applying Corpus Techniques to Climate Change Blogs. In A. Hardie and R. Love (eds.) *Corpus Linguistics 2013 Abstract Book*.

Andreas Schmidt, Ana Ivanova, and Mike S. 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.

Amelia Sharman. 2014. Mapping the climate change blogosphere. *Global Environmental Change* 26:159-170.

Rob Speer. 2016. ftfy software: https://ftfy.readthedocs.io/en/latest/

Xiaodan Song, Yun Chi, Koji Hino, and Belle L. Tseng. 2007. Identifying Opinion Leaders in the Blogosphere. *Procs. ACM CIKM '07*: 971-974.

Cass R. Sunstein. 2008. *Republic.com 2.0*. Princeton University Press.