Proceedings of the seventh Web as Corpus Workshop (WAC7)

Edited by Adam Kilgarriff and Serge Sharoff

Pre-WWW2012 Workshop, 17 April, 2012

Contents

Preface

> We want the Demon, you see, to extract from the dance of atoms only information that is genuine, like mathematical theorems, fashion magazines, blueprints, historical chronicles, or a recipe for ion crumpets, or how to clean and iron a suit of asbestos, and poetry too, and scientific advice, and almanacs, and calendars, and secret documents, and everything that ever appeared in any newspaper in the Universe, and telephone books of the future…

Stanisław Lem (1985). *The Cyberiad*, translated by Michael Kandel.

This year sees the seventh Web-as-Corpus workshop. Over that time we have explored a few parameters: adjoined-to-a-conference vs. free-standing; Europe-Africa-America; corpus linguistics venues and computational linguistics ones - but not, until now, web ones. If computational linguistics is the parent discipline for the group, then web research, as identified by the WWW conference, is the uncle who we don't see so often, but of whom we hear tales of travels and adventures and fortunes gained and lost, temples bedecked with gold and jewels but also dragons' dens, so it is with excitement and trepidation that we have booked a date with this wayward soul and anticipate getting to know him better. We look forward to the first WAC workshop to be held in conjunction with a WWW conference.

Adam Kilgarriff, Serge Sharoff
17 April, 2012, Lyon, France

Exploiting the Web for Text and Language Reuse Applications

Invited Talk

Benno Stein
Bauhaus-Universität Weimar
Bauhausstraße 11
99423 Weimar, Germany
benno.stein@uni-weimar.de

ABSTRACT

We will discuss backgrounds, technology, and applications developed in the Webis Research Group, whereas the talk's common thread is the exploitation of the web as a corpus. Three different applications will reveal different rationales and possibilities when operationalizing text reuse and language reuse on a large scale.

1. The Netspeak word search engine reuses the web as a corpus of writing examples. It indexes the web in the form of n-grams and implements a highly efficient wildcard search on the top. Netspeak supports writers by retrieving matching n-grams, which are ranked according to their occurrence frequency and which allow for judging a phrase's commonness compared to alternatives. The talk will highlight some of Netspeak's indexing concepts [1].

2. Query segmentation means to group the words of a search query into contiguous sequences without changing the word order; practically, it corresponds to the introduction of quotes indicating words that together form a concept. Query segmentation is receiving much attention since search engines are trying to guess concepts in queries automatically. The talk will outline the strategy of the currently best performing segmentation approach, which reuses the web as a corpus of frequently used phrases and concepts [2].

3. Assessing the effectiveness of plagiarism detectors is interesting and challenging [3]. The evaluation approach of the plagiarism detection competition PAN 12 will focus on the so-called candidate retrieval step: before two documents can be compared as to whether one contains a reused passage of the other, suited candidates need to be retrieved from the web. At PAN 12 the respective capabilities of the participating detection approaches will be judged by reusing a 25TByte large portion of the web, the ClueWeb09 corpus. The talk will introduce our strategy to organize such a competition.

[1] Benno Stein, Martin Potthast, and Martin Trenkmann. Retrieving Customary Web Language to Assist Writers. In Cathal Gurrin, Yulan He, Gabriella Kazai, Udo Kruschwitz, Suzanne Little, Thomas Roelleke, Stefan M. Rüger, and Keith van Rijsbergen, editors, *Advances in Information Retrieval. 32nd European Conference on Information Retrieval (ECIR 10)*, volume 5993 of *Lecture Notes in Computer Science*, pages 631–635, Berlin Heidelberg New York, 2010. Springer.

[2] Matthias Hagen, Martin Potthast, Benno Stein, and Christof Bräutigam. Query Segmentation Revisited. In Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M. P. Ravindra, Elisa Bertino, and Ravi Kumar, editors, *20th International Conference on World Wide Web (WWW 11)*, pages 97–106. ACM, March 2011.

[3] Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. An Evaluation Framework for Plagiarism Detection. In Chu-Ren Huang and Dan Jurafsky, editors, *23rd International Conference on Computational Linguistics (COLING 10)*, pages 997–1005, Stroudsburg, PA, USA, August 2010. Association for Computational Linguistics.

Understanding the composition of parallel corpora from the web

Marco Brunello
Centre for Translation Studies
University of Leeds
Leeds, LS2 9JT, United Kingdom
mlmb@leeds.ac.uk

ABSTRACT

Although it is fundamental to have a good fit between the text typology of training data and to-be-translated data in machine translation, there is a lack of studies on analysing parallel data under this point of view. This paper describes some studies made with the aim of understanding the composition of parallel corpora, in particular by using topic modeling.

Keywords

web as corpus, machine translation, parallel corpora, topic modeling

1. INTRODUCTION

Parallel texts have always been employed in translation studies, and in the last few decades large progresses in computational linguistics led the sub-field of machine translation to make wide use of large collection of aligned parallel corpora. Particularly those paradigms that strongly rely on the exploitation of this kind of resources to be used as training data, such as statistical machine translation (SMT), took advantage of this possibility. But, if the building of monolingual resources from the web comes with some undeniable issues - unknown overall composition of the Internet, continuous contents changing, unbalancedness towards particularly widespread languages or text varieties - the operation of retrieving multilingual data requires further technical skills for locating and pairing translated texts. As a consequence, the research in this field has not prospered in the same measure of monolingual web-as-corpus linguistics. Nevertheless, the use of the web as a source for parallel corpora is not an inert field of study. There is a quite rich tradition of studies in this direction, which includes a number of automatic tools that have been developed by several research groups, able to retrieve parallel data from the web (see Section 3.1).

However, it seems that not enough attention has been put on understanding what kind of documents are contained in parallel corpora in relation to their text variety. On one hand it is clear that, in a situation of scarceness of publicly available parallel corpora (especially when dealing with systems that base their successfulness on the employment of large collections of bitexts), the rule "more data is better data" is applied. On the other hand it could be still useful to have a clear knowledge and control of the kind of text to be translated using parallel data, especially in relation to their text variety, whatever is the chosen technology.

Even just the use of topic modeling techniques [14], employing an unsupervised method able to understand the composition of a corpus guessing what are the topic domains contained in a corpus, could be enough to give an idea of what are the kind of documents particularly interesting or useful for a particular task, such as the creation of translation memories to be used for fuzzy matches in a computer-assisted translation program or the selection of the most suitable training data for SMT. In this paper some exploratory studies conducted on two English-Italian parallel corpora are presented. In the first case, an already existing and very a well known resource, Europarl [7], has been used, while in the second case the employment of an automatic collection of translated texts from the web is described in order to collect data that will be part of a new parallel corpus. In this second case, the analysis will also help the reader to understand which topics are covered by the parallel web, and their potential usefulness in translation tasks.

The remainder of this document is organized as follows: Section 2 shows the experiments done on Europarl; Section 3, after showing previous related works on the collection of parallel corpora from the web, presents the way in which some of these techniques have been applied to collect part of a general-purpose English-Italian parallel web corpus, and the related topic modeling analysis; in Section 4 future directions of the results are suggested.

2. EXPERIMENTS ON EUROPARL

2.1 Topic modeling

As mentioned before, when talking about parallel corpora, Europarl is a very well known resource, widely employed by SMT scholars and researchers, since it provides aligned parallel texts up to 50 million words per language for the main European languages. Although it has been used in a variety of experiments, to our knowledge none of them has tried to explore its composition in terms of the arguments discussed in it. Being its parallel texts transcriptions of proceedings of the European parliament, the supposition that the communicative situations are repetitive, both in terms of register of the utterances and of content of the communications, could be fairly reasonable.

In order to verify this hypothesis, a topic modeling on the English side of the English-Italian Europarl has been performed. As the choice of the topics is arbitrary, different customisations have been tried, finally deciding for a number of 20 topics with hyperparameter optimization. Results are shown in Table 1.

It is clearly possible to associate some of these clusters of words to specific topics. For example Topic 1 is related to energy production (and related environmental issues); 3 is about warfare (with a stress to the Iraq war); 4 immigration; 5 medical treatment of addictions; 6 food production; 7 Euro-economics; 8 global market; 9 Middle East; 12 transports; 13 finance; 17 human rights; 18 Eastern Europe; 19 primary sector; 20 legal aspects. The remaining groups of words appear to be quite similar, since they do not suggest very specific terminology and contain almost completely terms around the European parliament activities, widely used throughout the debates in the European parliament: *president*, *report*, *parliament*, *council* etc.

Taking a look at the distribution of the single document across the topics, it appears that very few documents show a remarkable unbalancedness towards the first topic they belong to, meaning that the affiliation of a single document to specific topics is weak. Probably this is due to the fact that sessions of the European parliament can deal not only with a single topic at time, and it is not the case where a corpus document always corresponds to a single argument. It can be concluded that, although this experiment has led to a better understand of the composition of Europarl (or at least of the portion of English Europarl aligned with its Italian counterpart), it has demonstrated that this corpus is too homogeneous to be used for the extraction of subsets for SMT.

2.2 Document similarity experiment

Taking into account the conclusion achieved in the previous section, an attempt of selecting a subset of training data for a specific translation task from Europarl has been made anyway. The aim is to pick up a selected portion of Europarl documents that are recognized to be the most similar to those to be translated, and use them as training data for SMT. Since the topic modeling has proved not to be very helpful for this purpose, the cosine similarity measure has been chosen as alternative to compute the distance between a test document to be machine-translated and all the documents in Europarl. The test document has been randomly selected from the Internet, it is a journal article about a controversy in the Catholic church, dated 10 September 2009; it was written in Italian and provided with English human translation (which will be used later as benchmark for MT evaluation); its size is around 1350 words per language on 47 sentence pairs.

In order to select and use the most similar data in Europarl to this document the following procedure has been followed:

Table 1: Topics in Europarl.

Topic	Keywords
1	energy climate european change emissions eu environmental gas nuclear industry policy renewable research europe countries environment states sources development
2	european europe president union mr people parliament citizens treaty political today presidency time eu constitution member states future world
3	international people peace situation war aid united union military european resolution president security government support humanitarian mr country iraq
4	european states member rights report data protection eu legal justice terrorism immigration law asylum citizens countries people union crime
5	health research people programme european diseases tobacco drugs states disease member human public europe information patients care framework treatment
6	directive food products health environment safety proposal animal protection animals amendments consumer legislation waste water environmental consumers commission substances
7	economic euro financial european growth monetary bank stability crisis economy policy central states market currency pact countries markets investment
8	countries trade development world eu european agreement union developing economic international africa wto china aid cooperation negotiations agreements global
9	israel education european palestinian cultural people culture programme sport israeli young peace languages middle support palestinians europe media training
10	council european union policy countries presidency mr president rights parliament political states office common agreement security enlargement summit process
11	commission mr president member european important time make commissioner made work states debate point parliament question council fact clear
12	transport safety european road air proposal tourism traffic rail report maritime directive europe sector mr sea environmental states passengers
13	budget commission parliament financial european funds committee programme policy year eur budgetary money report support council fund aid court
14	parliament mr vote report president committee amendment group european members procedure amendments house minutes rules mrs resolution voting rapporteur
15	report social women european policy eu states member employment people development work support voted europe economic writing union regions
16	report european union mr parliament policy council countries committee social economic europe community people president agreement treaty question employment
17	rights human people country president democracy political situation resolution government freedom china death democratic european eu mr respect elections
18	eu european union turkey russia countries accession country ukraine russian negotiations turkish political relations enlargement romania report cooperation region
19	fisheries fishing agricultural policy sector report production proposal farmers support agriculture rural market measures european aid regions commission reform
20	directive market services european proposal report member competition states parliament internal companies legal workers public rights legislation regulation protection

1. The cosine similarity has been computed between the test doc and each one of the 6,216 Europarl files (on the English side of the EN-IT subcorpus);

2. Files have been sorted according to the cosine similarity score and the first 500 have been selected;

3. The result of this selection (303,615 sentences pairs, around 7 million words per language) has been used to train a SMT system in Moses [8];

4. The obtained parameter file has been used to translate the test document (in the English>Italian direction);

5. The previous steps 3 and 4 have been repeated substituting the training data selected with cosine similarity with other 500 documents randomply extracted from Europarl, in order to obtain a term of comparison and validate the quality of the resulted translations;

6. MT evaluation of the quality of these translations has been carried out by using BLEU [12];

7. The whole above described process has been repeated in the opposite language direction (Italian>English).

The results of this analysis are presented in Table 2. They confirmed that the selection of a subset of documents that are more similar to the to-be-translated document is useful, giving - at least according to the automatic evaluation conducted by BLEU - a better translation than using random documents even in a corpus like Europarl that is circumscribed to few communicative situations. However, there are some considerations that need to be done about these results: although they are positive they are not extremely exciting, especially considering that the random training set was much smaller than the one selected with cosine similarity (117,973 sentences pairs and about 2 million words per language). This is most probably due to the fact that cosine similarity calculates the similarity between the two documents as a single whole feature, without distinguishing between the several aspects that make a texts similar to another one, like terminology, sentence structure, grammar, length etc. So there is the possibility that training data appropriate and useful in terms of genre or domain but with size is not good to test texts are discarded from this kind of selection.

To sum up, some strategies to better understand the composition of one of the most used parallel resources freely available to the MT community has been applied, including the possibility to select the most suitable data for a specific SMT task among all the documents contained in it. The conclusion is that, even if these strategies can be helpful for their usability on Europarl, they are quite limited for the intrinsic features of this corpus - that provides a huge quantity of data but without a very big assortment of textual varieties. Previous research about domain adaptation applied to SMT [9] has shown how some help could come from integrating a benchmark resource like Europarl with other data, obtained by retrieving parallel corpora from the web. Europarl itself could be considered as a webcorpus, since their authors have built it after having downloaded and aligned the proceedings from the website of the European parliament[1], despite the fact that these texts were not originally created with the specific purpose of being published on webpages. In any case, in the next sections the way to explore this possibility is described, practically considering which instruments can be used to retrieve parallel documents from the web and understand their nature.

Table 2: BLEU score for the cosine similarity experiment

Direction	Training set	BLEU score
IT>EN	500 most similar	27.5
IT>EN	500 random	26.1
EN>IT	500 most similar	26.5
EN>IT	500 random	23.3

3. PARALLEL CORPORA FROM THE WEB

Crawling the web in order to find textual data is a common practice in corpus linguistics, and even if coming with a surplus of difficulties comparing to monolingual corpus creation (finding parallel pages or websites or pairing a webpage with their translated counterpart in another language are non-trivial tasks) also the creation of parallel corpora from the web was explored especially in the last decade. However, the literature shows that these technical difficulties led the researchers in the sector to focus more on the mechanisms needed to create parallel webcorpora rather than on an in-depth analysis of the results from the point of view of the kind of texts that is possible to find on the Internet. In the next few subsections an outline of the background studies on this topic will be given, followed by the way in which some of these strategies have been applied in the present research in order to understand the composition of a particular region of the parallel web.

3.1 Related works and known issues

The forerunner of using the web as a source for collecting parallel corpora is Philip Resnik and his sistem STRAND, described since the late nineties in a series of papers. In the latest, *The web as parallel corpus* [13], the core STRAND system is explained as well with several improvements comparing to previous versions. The main idea behind this approach is to find web pages that exhibit a parallel structure at the level of url and/or page composition, and that could be mutual translations; in the practice this is done relying on the performances of AltaVista advanced search engine options that permit to find pages containing hypertexts links to different language versions of the same document (parents) or pages that contains a link to a translation of their content in another language (siblings). The so-retrieved pages are then subject to a candidate pairs detection task that can be carried out with several strategies, like automatic language identification, url matching, document lengths and in the last version also a content-based similarity measure, to detect pairs of pages that do not present similarity just at the level of structure.

As shown with STRAND, the operation of grabbing parallel texts from the web is usually divided into two steps: loca-

[1] http://www.europarl.europa.eu

tion of websites that may have translated texts, and extraction and alignment of bitexts from these candidates. Other systems, developed independently from STRAND but employing similar approaches, have seen the birth in the same period, and some of them have explicated some useful (although not universally true) presuppositions that can be made when starting a collection of parallel texts from the web, like assuming that parallel texts usually are present in the same site [4] and that national top level domains are expected to have sites in the language of respective countries [10].

Several approaches in literature rely on the functionalities given by commercial search engines, like Altavista for STRAND or the application programming interfaces (APIs) to commercial search engines Google [11] or Yahoo! [1]. It is worth to point out -for reasons explained in the next sections - that the reliability on these systems, provided by famous search engines, comes with some disadvantages: there are intrinsic limitations with regards to the number of results per query and the unknown criteria about how documents are selected, their availability is for undefined periods of times after which the service may be no longer supported or available to users[2]. However, even if coming with these disadvantages the reliability on search engine functionalities is a possibility that gives undeniable advantages, considering the fact that it gives easy access to previously unknown parallel texts as the next section will explain in detail.

3.2 Corpus construction

The strategy here used to mine the web looking for previously unknown parallel pages is very similar to the original one described by Resnik: to make use of a search engine relying on its specific research functionalities. Resnik used Altavista, but the employment of these strategies has the drawback of relying on possibly discontinued services. In fact at the moment it is not possible to exactly replicate their specific algorithm because this search engine is no more available with the same options[3] of the time that the article was written.

However, it has been possible to re-implement a similar procedure using the search engine query algorithm that is part of the BootCaT toolkit, as just said currently relying on Bing. As seeds tuples the ones produced to build the large web corpus of English ukWaC [6] have been used, adding to each of these 1000 lines the use of two advanced operators: `site:` and `inanchor:`. The first is used to look for sites that fall under the intended national top level domain (in this case `.it`) and the second to find pages containing a specified term in the anchor text, in this case common English-related features of URLs (`en`, `eng`, `english`).

In practice, the search has been for pages in Italian websites that most likely contain English versions of their content.

Table 3: First 10 lines of the seeds list.

```
inanchor:en site:it grey gently
inanchor:en site:it drawing totally
inanchor:en site:it path eating
inanchor:en site:it watching explanation
inanchor:en site:it dealt lack
inanchor:en site:it radical organised
inanchor:en site:it relationships studied
inanchor:en site:it gets accused
inanchor:en site:it conservative hoping
inanchor:en site:it realise increasing
```

Queries are issued asking for 50 results per query (this is the maximum available from Bing APIs).

Table 4: Results for the first 3 queries on an English-Italian pair.

```
CURRENT_QUERY inanchor:en site:it grey gently
http://www.domusweb.it/en/architecture/teshima-art-museum-/
http://gilda.it/gandalf/italiano/giochi_di_ruolo/girsa_rolemaster/moduli/
amroth/amroth.htm
http://www.beppegrillo.it/en/politics/
CURRENT_QUERY inanchor:en site:it drawing totally
http://en.metals.it/productive-cycle-punching-c-104_133.html
http://flashandpartners.it/en/
http://www.dieproofs.it/english/prove_artista_eng.html
http://digilander.libero.it/cuoccimix/ENGLISH-automotorusse4(lada).htm
http://www.digicult.it/digimag/article.asp?id=1141
http://www.domusweb.it/en/architecture/post-carbon-loft-
http://architettura.it/artland/20020515/index_en.htm
http://en.museiincomuneroma.it/mostre_ed_eventi/mostre
http://www.beppegrillo.it/en/2010/06/
http://www.beppegrillo.it/en/information/
http://www.asianews.it/index.php?l=en&art=22711&size=A
http://www.domusweb.it/en/products/?idtema=5515?idtema=5530&inizio=25&da=1
http://www.domusweb.it/en/products/?idtema=5515?idtema=5562&inizio=13&da=1
http://www.disabilitaincifre.it/allegati/RECOMMENDATION_R(92)6.htm
http://www.beppegrillo.it/en/2009/08/
http://www.domusweb.it/en/products/?idtema=5515?idtema=5322&inizio=49&da=1
http://archivio.lanottebianca.it/nb2006/en_programma.html
http://www.pierpaoloricci.it/download/downloadsoftware_eng.htm
http://www.beppegrillo.it/en/politics/
http://archivio.lanottebianca.it/nb2005/en_programma.html
http://www.beppegrillo.it/eng/politics/
CURRENT_QUERY inanchor:en site:it path eating
http://www.guidatoscana.it/en/massa-carrara/visitare-massa.asp
http://www.visittrentino.it/en/localita/lavarone
http://www.ananda.it/en/courses/courses-meditation-and-self-realization
http://www.holly-wood.it/mlcad/install-en.html
http://www.visittrentino.it/en/vacanze_a_tema/neve/ski_area/dett/
ski-area-pampeago-predazzo-obereggen?areaId=A10
http://www.italia.it/en/discover-italy/emilia-romagna/ferrara.html
http://www.italia.it/en/discover-italy/tuscany/florence.html
http://www.beppegrillo.it/eng/ecology/
http://archivio.lanottebianca.it/nb2005/en_programma.html
http://www.beppegrillo.it/en/2009/08/
http://www.beppegrillo.it/eng/politics/
```

At the end three lists of urls for each language, sorted and unified in order to have one single list per language without repetitions have been produced. In order to extract those pages that actually are English translations of other Italian webpages on the same website, the final url list has been semi-automatically processed.

At this point such pages have been downloaded, and analysed by using Mallet. Results of this topic modeling are shown in table 5.

3.3 Analysis of results

One thing that need to be clarified now is that this is not a definitive description of the overall composition of the English-Italian portion of the Internet: the following analysis shows what is possible to intuitively retrieve using the search engine method and have an approximate idea of the kind of the indexed websites that represent a bilingual web space. However, since the accessibility to single site depends on their presence and positioning on the search engines, the

[2] BootCaT [3] used Google APIs in its original version, but then moved to Yahoo! after Google started giving strong limitations to its service and now BootCaT relies on Bing APIs after Yahoo! discontinued the use of their SOAP APIs (see `http://bootcat.sslmit.unibo.it/wiki/doku.php?id=release_notes:frontend:0.60`).

[3] The shut down of Altavista by its owner Yahoo! began in May 2011.

Table 5: Topics in the EN-IT corpus

Topic	Keywords
1	music art design work fashion world italian time film years style architecture de project works york great life la
2	di la engine il px car de brutale che pic version del mv agusta pics hp della news con
3	january inter milan time news photos derby don team ac great good people day back comments ll search italia
4	china peace arab christians vatican chinese world muslims christian government years year country case rsquo bishops samir mgr people
5	hotel di area sea wine day city km offers visit rooms room florence centre located italy free beautiful town
6	century di museum city church ancient san town built art roman del area rome st building palazzo history archaeological
7	italian international italy information university research di european law services public development students data company people financial system management
8	water high production system products quality made product time light energy range type materials oil process air design control
9	data time set file software information user version click web function download find system show distribution site number image
10	church god pope life world faith vatican people time Catholic holy benedict xvi christ great cardinal council human ii

present analysis shows what is actually possible to retrieve with this method; other existing strategies to find previously unknown multilingual websites [2] rely on the crawling of large web directories like the Open Directory Project[4], and again the possibility of benefiting from parallel sites is circumscribed only to those sites that are listed (in this case by human users to specific directories).

Considering the generated topic: most of the documents shows a strong belonging to the first topic they have been associated with, so we can graphically generate a plausible representation of the distribution of each topic in the whole context of the corpus.

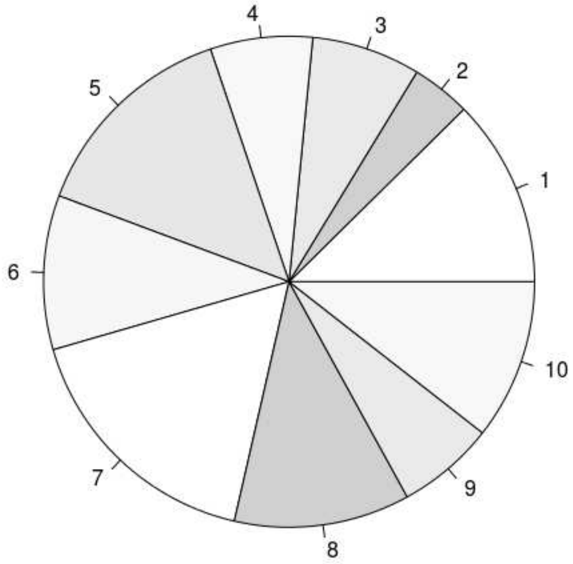

Figure 1: Distribution of the topics

A sample of 10 randomly-selected documents for each of these topic have been selected and read by the author in order to understand the overall nature of the single portions of the corpus associated to each group of keywords.

1. The first topic suggests pages containing text related to fashion, art exibitions, recent trends in architecture, cinema, music etc. In fact it appears to be composed by articles coming from (but not exclusively) the online version a famous fashion and lifestyle magazine.

2. The second topic appears to be instead more problematic to be understood, since it contains several stopwords in Italian, such as article and prepositions; it is worth to say that English stopwords have been stripped out when preparing the corpus for topic modeling, since this is actually an English corpus. The generation of this topic can suggest the deletion of the original language counterpart's stopwords as well, but this can be risky since it would modify several compound proper nouns contained in the corpus as it has been revealed in the case. This portion of the corpus collects pages mainly from two sites: a fan site about vintage cars produced in the Soviet union, and a motorcycle manufacturer site; the unknown words *agusta* e *brutale* have reference to proper nouns of brands.

3. The third topic contains words related to football, in fact great part of this portion is made by news about football matches. However these keywords could be misleading since this part of the corpus also contains news about only other sports and other kinds of entertainment, such as tv programs and videogames. Words like *comments*, *back*, *search* are due to the presence of recurring boilerplate in the football news, that - again - almost completely have reference to a single website.

4. Also the documents collected around the fourth topic have reference to a single website, in this case a news website promoted by a Catholic missionary group (that generated a list of religious-related terms).

[4] `http://www.dmoz.org/`

5. The fifth topic instead collects contributions from many single sites. Clearly suggested by the keywords, the manual analysis confirmed the presence of English versions of webistes of tourist accommodations such as hotels, residences, bed & breakfast etc.

6. The sixth topic is related to texts about cultural tourism and history of art. The samples have shown that websites of artistic attractions and personal pages with more pedagogical purposes belong to this category.

7. The seventh topic is less homogeneous that the previous ones. Contributes to this large part of the corpus come from presentation of university programmes and related bureaucratic guidelines, but also legal disclaimers from private companies and translations of legal statuses.

8. This is another topic that collects contributions from many sources but present a quite specific genre of texts: companies presenting their products. Words like *high*, *quality*, *production*, *range* suggests the intention of promoting themselves, and this was confirmed by the sample data. The little portion manually analysed revealed the presence of several kind of companies, from the primary sector to services; but the biggest part appeared to be related to mechanical and electrical product engineering.

9. The ninth topic appears to be about computing. Around this topic are collected webpages of various origins, from academic workshops to web tutorial to description of programs, plugins etc.

10. As suggested by the keywords, the tenth topic collects articles around it articles about the Catholic church, in particular news and debates. Most of the documents have reference to two online magazines, but in this portion of the corpus there are also pages coming from other sources and not necessarily related to the Catholic church hot topics, since they discuss about philosophy, sacred music or other religions.[5]

This analysis has provided an overview of the downloaded documents, and the possible composition of the so generated corpus. The topic modeling technique revealed the major presence of documents taken from recurring websites, in a way that several topics corresponds to particular webistes and their content. Even if this circumscribes the possibilities of having variety in the corpus, it can give advantages for the problem of retrieving document pairs, since it allows to develop a specific strategy for these large websites basing on how they organize their content. On the other hand, there are groups of different websites but having similar purposes, like showcasing products and services or displaying guidelines and rules. In this case there are recurring language structures spread across pages coming from different sources, that could be useful e.g. for generalizations about grammar or terminology.

4. FURTHER WORKS AND CONCLUSIONS

Since the purpose of this paper is only to describe a strategy to explore and understand the composition of a bilingual web space, only an aspect of the building of a parallel corpus from the web (location of webpages that may contain parallel texts) has been taken into account, i.e. the ability to find pages that contain translated text[6]. This is just the first step, as it needs to be followed by further fundamental operations, above all the location of the counterparts in other(s) language(s) of the previously obtained webpages and the sentence alignment of their content.

This study is part of a major PhD project aimed at exploring how much the successfulness of SMT technology depends on the exploitation parallel corpora basing on a good match between the text typology (genre and domain) of training data and to-be-translated texts. This means that the preliminary study here presented is going to be continued with the creation of some parallel corpora from the web following the strategy described in section 3.2, and the mentioned further steps. Previous literature has shown that the generation of candidate pairs can be performed via a series of heuristics such as url substitution rules, analysis of document lengths and structural filtering. Since there is not an established state-of-the-art system to perform this job and very few tools able to do that are freely available[7] - some of them are currently experimented, and possibly integrated in a single framework that would be able to cover the whole chain, from the search for parallel pages to the extraction of parallel pairs and their alignment. Another thing that has not been mentioned in this paper but that can be very useful to expand the size of a parallel corpus is that the experiment here described deals with single pages rather than whole websites: even if the search engine selected only a webpage from a site, there is the real possibility to find more parallel texts, going up to the main website and exploring the tree of its contents, collecting all the others parallel pages contained in it. To conclude, this was an example focused on a particular language pair (Italian-English), but it would be interesting to explore different language pairs in order to understand what different kind of parallel data are on the web depending on particular language pairs.

To sum up, in this paper some studies about the importance of the analysis of parallel corpora have been conducted, starting from the considerations that not only it is advisable to have knowledge of the data used (in particular, in the second part of the paper, the exploration of the composition of a parallel corpus retrieved from the web has been attempted), but also that this can be helpful in order to select the most suitable textual data for our specific purposes. The strategies here proposed can have wide application for everybody who wants to better benefit from existing or to-be-constructed parallel resources, from the creation of translation memories for computer-assisted translation to the creation of larger parallel corpora. The next step will be that of exploring this last possibility, since there are not standards about the dimensions of parallel corpora from the web and their size can remarkably change among different language

[5] The test article described in section 2.2 was taken from this collection.

[6] Scripts and commands employed in the experiments here described can be found at `http://smlc09.leeds.ac.uk/marco/tools.html`.

[7] A rare example is Bitextor [5].

pairs.

5. ACKNOWLEDGMENTS

The research described in this paper is supported by FP7 ICT grant
ACCURAT Grant agreement no.: 248347

6. REFERENCES

[1] J. J. a. Almeida and A. Simões. Automatic Parallel Corpora and Bilingual Terminology extraction from Parallel WebSites. 2010.

[2] L. Barbosa, S. Bangalore, and V. K. S. Rangarajan. *Crawling Back and Forth: Using Back and Out Links to Locate Bilingual Sites*. 2011.

[3] M. Baroni and S. Bernardini. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of the LREC 2004 conference*, volume 4, pages 1313–1316. ELRA, 2004.

[4] J. Chen and J.-Y. Nie. *Parallel Web text mining for cross-language information retrieval*, pages 62–77. Paris, 2000.

[5] M. Esplà-Gomis and M. L. Forcada. Combining content-based and URL-based heuristics to harvest aligned bitexts from multilingual sites with Bitextor. In *Fourth Machine Translation Marathon Open Source Tools for Machine Translation*, 2010.

[6] A. Ferraresi. Building a very large corpus of English obtained by Web crawling: ukWaC, 2007.

[7] P. Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, 2005. AAMT.

[8] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[9] P. Koehn and J. Schroeder. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 224–227, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[10] X. Ma and M. Liberman. *Bits: A method for bilingual text search over the Web*. 1999.

[11] M. Mohler and R. Mihalcea. Babylon Parallel Text Builder: Gathering Parallel Texts for Low-Density Languages. In *LREC'08*, 2008.

[12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[13] P. Resnik and N. A. Smith. The Web as a parallel corpus. *Comput. Linguist.*, 29(3):349–380, Sept. 2003.

[14] M. Steyvers and T. Griffiths. Probabilistic Topic Models. In T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning.* Laurence Erlbaum, 2006.

A Corpus of Online Discussions for Research into Linguistic Memes

Dayne Freitag
SRI International
freitag@ai.sri.com

Ed Chow
SRI International
edchow@ai.sri.com

Paul Kalmar
SRI International
kalmar@ai.sri.com

Tulay Muezzinoglu
SRI International
tulay@ai.sri.com

John Niekrasz
SRI International
niekrasz@ai.sri.com

ABSTRACT

We describe a 460-million word corpus of online discussions. The data are collected from public news websites and community-of-interest Internet forums, and are designed to support research on the propagation of socially relevant ideas, a.k.a., "memes." A structural and statistical description of the corpus is given, and the employed methods of website monitoring, collection, and extraction are described. We also present preliminary linguistic research on the corpus. We show that the corpus represents language from a wide variety of social and psychological communities, that discussion structure and popularity can be predicted in large part from lexical analysis, and that standard epidemiological models provide good fit for diachronic patterns of population-level lexical adoption.

Categories and Subject Descriptors

H.2.4 [**Database Management Systems**]: Textual databases; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*

General Terms

Corpus analytics, Memetics, Information diffusion

1. INTRODUCTION

Over the relatively short period since its inception, the Web has assumed an increasingly central role in the dissemination of information and the spread of ideas. The widespread adoption of social media, an even more recent phenomenon, has dramatically decreased the friction with which both trivial and momentous ideas spread. In the past, these socially relevant ideas, these *memes*, might have gained most of their force through official promulgation. In the modern landscape, there is a much stronger bottom-up component to this spread, resulting in more turbulent and, arguably, more interesting patterns in the diffusion of influence.

Information diffusion research has attempted to reconstruct and model these influence patterns, exploiting the computational accessibility of both the linguistic (tweets, blogs, wall posts) and social (friend or follower networks) dimensions of social media. Much of the early work focused on the relatively professional utterances of bloggers, seeking to recover the transmission trajectories of memes (typically URLs or phrases), and to quantify influence as a feature of individual blogs or bloggers [6, 9]. An epidemiological analogy is often applied to the spread of memes through such networks, and models derived or borrowed from this analogy have shown some success in accounting for observed patterns [2, 5]. With the increasing popularity of microblogging, studies of information diffusion in platforms such as Facebook and Twitter have yielded insight into idea propagation and social network formation closer to the "grass roots" [13, 10].

The work described in this paper continues this trend of research away from the professional pundit toward the average citizen, with an emphasis on online discussions. As a rule with a few notable exceptions, Web sites providing news and commentary include comment boards where the reader can respond to specific articles or to other commenters. Such discussions can also be found on special interest Web sites, what we call "communities of interest" (COIs). As will become clear, the harvesting of such discussions poses challenges that have impeded their widespread use in information diffusion and computational linguistics research. But as we also attempt to show, online discussions promise novel socio-linguistic insights.

Our "meme epidemiology" project pursues insights suggested by the epidemiology analogy, attempting to elaborate it in several ways. First, we are interested in what can be learned by treating discussions as "outbreaks." Like disease outbreaks, discussions grow over time and at different rates. Some become truly huge, while many fail to develop at all. We seek to identify the factors that produce these differences, exploiting the overt connection between an article and the discussion it engenders. Second, we assume that, just as in epidemiology, the energy a discussion or idea exhibits is at least partly attributable to the community or population that hosts it—an assumption that our collection of COI data allows us to test. Finally, we believe that conventional epidemiological models are much more directly applicable to idea diffusion than existing research might suggest. Later, we show that an SIR compartmental model, borrowed with few modifications from epidemiology, accurately models the temporal distribution of lexical expression patterns over several years.

In this paper, we present a 460 million-word corpus of online discussions. We begin in Section 2 by describing the corpus contents and data model. We develop a common vocabulary for corpus elements, and we provide an overview that reveals some of its salient statistical properties. In Section 3, we present a purpose-built data collection system that has been used to monitor, collect, and extract the data from multiple sources. In particular, our presentation highlights some of the challenges we have encountered in the design of this system and the maintenance of a coherent corpus. Finally, Section 4 presents initial results from three areas of linguistic research being conducted using the corpus: (1) modeling and prediction of discussion structure, (2) linguistic variation between and within website communities, (3) and meme propagation.

2. CORPUS DESCRIPTION

We have been collecting the corpus that is the focus of this paper for nearly a year from public sources. Major features of the collection system and database structure have been stable for approximately six months. We continue to collect data from the sites listed below, and to add to the list of sites.

2.1 Data model

We harvested from 24 distinct websites, each of which have their own way of providing users with the ability to conduct discussions online. As a result, the type and organization of data present on each site may be different from one site to the next. It is therefore necessary as a first step in developing and studying the corpus to develop a common representation for all the linguistic data present—one which generalizes well across the multiple sites and allows for discussion and analysis across the entire corpus.

The data model is centered on *discussions* as the main representational unit. Discussions are built from two types of discourse unit—*articles*, which we use to refer to an initial posting of some content (typically a news article or editorial) and *comments*, which refer to any subsequent statements made in response. Each comment also has an *attachment* relation linking it either to another comment (when one commenter replies to another) or directly to the initiating article (we refer to this latter type as *root attachment*). All comments and articles are assigned a *posting date* (which may include time-of-day information if it is available). Each comment's *author* is also obtained, using public user handles when available. The authorship of news articles, in contrast, is not currently available, as we do not have a sufficiently robust mechanism for extracting this information from its embedded position within article text.

We distinguish two main types of websites—*community-of-interest (COI) forums* and *news sites*—each type providing certain advantages of interest to the project. COIs explicitly group discussants into more or less culturally homogeneous populations, while news sites make explicit the connection between discussions and the real-world events to which they respond. The two types are distinguished primarily by the way that discussions are initiated (and by whom). For news sites, discussions are initiated by the posting of news articles or editorials that are written by professional authors who are typically not participants in subsequent discussion. Forum discussions, on the other hand, are initiated by discussants themselves, which means the "articles" are usually better described as a discussion "prompt" (though professionally-written articles, or hyperlinks to them, are sometimes posted as articles in forums). News sites and COI forums are also typically distinguished by the nature of their participant community. As the name suggests, COI forums have a more targeted set of common interests, and therefore draw a more focused set of participants.

We are interested in modeling discussions as linguistic objects in their own right, particularly the reply or attachment structure they display, but websites in the corpus often limit certain types of attachment, thus constraining the set of possible discussion threading structures. For example, some sites allow new comments to attach only to the most recently posted comments. In other cases, the recursive depth of the attachment tree is limited. Some web sites eliminate structure altogether, and do not allow comments to attach to other comments at all. These differences limit our ability to generalize some of our findings about discussion structure, but also provide an opportunity to learn about how such constraints affect information propagation. Nonetheless, the applicability of the data model just described is not affected by these differences.

2.2 Descriptive statistics

The corpus consists of approximately 460 million words extracted from 24 websites.[1] A list of the collected websites is shown in Table 1, with those allowing for comment–comment attachment marked with an asterisk ($*$). As described in the previous section, it is useful to classify the sites into two main types: news sites and community-of-interest (COI) forums. In our selection of COIs, we are interested in choosing sites with a pronounced point of view, while sampling from as broad a range of persuasions as possible. Table 2 presents summary statistics for each of these two components of the corpus. The data show that comments tend to be longer in COI forums, and that COI forum communities tend to be smaller. Also note that the posting of articles is typical of news sites but not COI forums, though there are some exceptions to this (the COIs `richarddawkins.net` and `vanguardnewsnetwork.com` contain posted articles, and some news sites have a few discussions without a posted article).

For many of the websites (typically the news sites), historical data are not made publicly available, so the corpus only contains articles and posts from the period of the collection effort. This means that our archives of such sites contain data spanning periods between 3 and 6 months (depending on when the site was introduced to the collection queue.) Some sites (typically the COI forums) do provide this historical data. For these sites, the collected data spans periods ranging from 1 to 7 years. The website `animalsuffering.com` has the longest archive and contains data going back to 2004.

An analysis of the distribution of discussion size (i.e., the number of comments in a discussion) reveals interesting properties of the corpus. Namely, we studied the relationship between discussion size and discussion size frequency. In related internet phenomena such as social network connectivity, popularity of websites, or number of email contacts, it has been found that these distributions follow a power-law distribution [1]. But contrary to this pattern, we find that our data require a *sub*-logarithmic transformation of the two variables (discussion size and discussion size frequency) to produce a linear relationship. This suggests that the stochastic processes that are thought to underly some power-law distributions, such as preferential attachment, may not apply in a straightforward manner to our data [8]. We discuss this further in Section 4.

The data also reveal that temporal factors vary widely across sites. Discussions on the news sites `latimes.com` and `wsj.com`, for example, tend to dissipate rapidly, with 95% of comments occuring

[1] Collection of the corpus is an ongoing effort. Articles and comments continue to be autonomously collected, and websites are still being added to the collection effort. The description in this paper therefore applies to the state of the corpus as of January 2012, which represents about 6 months of data collection processing.

Table 1: A list of collected websites. Those allowing comment–comment attachment are labelled with an asterisk (*).

News sites	COI forums
bostonglobe.com	animalrightsdiscussion.com
foxnews.com*	animalsuffering.com
huffingtonpost.com*	boston.com
lasvegassun.com	conservativesforum.com
latimes.com*	hindudharmaforums.com
miamiherald.com*	kongregate.com
motherjones.com*	mothering.com
npr.org	mpacuk.org
nymag.com*	richarddawkins.net
reuters.com	thehighroad.org
washingtonpost.com*	vanguardnewsnetwork.com
wsj.com*	vegansoapbox.com

Table 2: Summary statistics for the two main components of the corpus: news websites and community-of-interest discussion forums.

	News sites	COI forums
# of websites	12	12
# of discussions	148,948	88,551
# of articles	116,449	13,842
# of comments	6,373,186	1,367,586
# of words in articles	53,241,204	6,965,108
# of words in comments	255,267,240	145,414,708
mean words per comment	40	106
mean words per article	457	503
mean unique commenters per site	26,525	5496

within 3 and 4 days, respectively, of the posting of an article. On COI forums, however, discussions have a longer life, with the same statistic for boston.com and mothering.com being 8 and 26 days respectively. Interestingly, however, we find that data from all of our sites fit well with a *log-normal* temporal distribution for comments posted in a discussion, an observation that matches findings in other dynamic processes on the internet, such as the evolution of internet meme popularity [3]. Daily, weekly, and seasonal variations in activity are also readily apparent.

3. METHODS OF DATA COLLECTION

Almost all news sites engage their readers by allowing comments to be attached to news articles. In fact, commenting has become so essential that there are now hosted services such as *Disqus*[2] and *Echo*[3] offering a comments platform. However, most of these web applications uses AJAX technology and require user interaction, making it very difficult to crawl such data [11].

Forum sites that create communities around a specific topic have long been around. Platforms used by such sites are more or less similar. While forum sites can be crawled using classical methods and do not use AJAX, they tend to require registration in order to access forum content.

Our data collection system shown in Figure 1 consists of a *discovery module* and an *extraction module*. It is designed to satisfy two major goals: to discover new article URLs from sites of interest, and to extract individual comments and articles. System capabilities include programmatic login and using the Tor[4] network for anonymity.

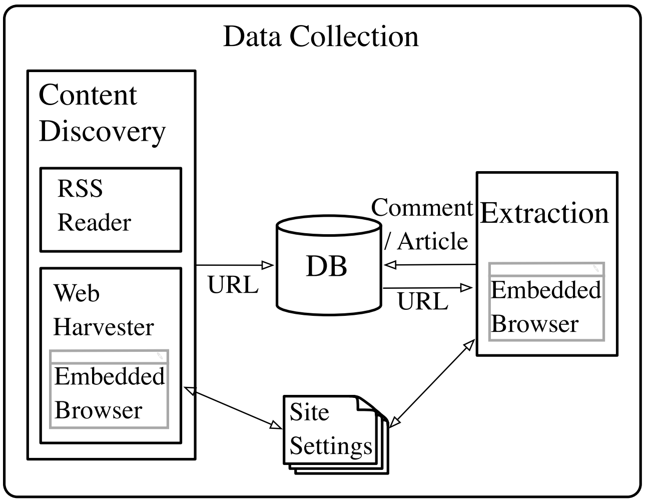

Figure 1: Data collection system

All twelve of the news sites we harvested, and three COI sites, provide RSS feeds. For these sites, an RSS reader probes feeds for new content, and stores the URLs in a database collection. For sites lacking RSS feeds, or to gather archival data, the process is driven by site-specific configuration files containing seed URLs. A web harvester with an embedded browser starts with these seed URLs, extracts all links and either stores them in a database or queues them and continues navigation. Discovered URLs are then picked up by the extraction module.

The extraction of the actual articles and comments from the HTML pages is the most challenging part: first, AJAX-enabled sites require user interaction with pages to initiate data requests before comments can be navigated; second, each site serves different metadata for comments, preventing the development of a unified data model. We address the first issue with browser-based harvesting, and the second issue with guided extraction and a schema-less document storage.

A generic configuration file used by our system is shown in Figure 2. It consists of sections with key-value pairs, usually expressed in JSON format.

While the *meta* and *login* settings are obeyed by both the harvesting and extraction modules, the *url-patterns* section is mainly used by the harvester. To limit the URL search space, only navigational URLs, such as pagination links, are followed. Links that match article patterns actually point to a main news article or to the head of a thread, and therefore are stored in the database for further processing. The *requestRate* in the *meta* section defines the delay between consecutive page requests from a single site; its default value is 15 seconds.

The remaining configuration sections are relevant to the extraction

[2] http://disqus.com

[3] http://aboutecho.com

[4] https://www.torproject.org/

```
[meta]
id          = SiteID
seeds       = {"urls":[seed1,seed2]}
tor         = 1
requestRate = 25

[login]
username    = {"by":"name","value":"username",
               "send":"myusername"}
password    = {"by":"name","value":"password",
               "send":"mypasswd"}
url         = http://login.url

[url-patterns]
navigation  = {"urls":[regexp1,regexp2]}
article     = {"urls":[regexp3],"save":"1"}

[articles]
showmore    = //a[text()='Single Page']

[article]
title       = //meta[@property='og:title']/@content

[comments]
root        = //tr[starts-with(@id,'CommentKey:')]
navigation  = {"xpath":"//div[@id='Paginator']
               //a[text()='Next']","loadfirst":1}
commentsLink = //td[@id='CommentsHead']/a

[comment]
author      = .//td[@id='profile']/a
replyTo     = .//a[@class='reply-link']/@href
```

Figure 2: Sample configuration for navigation and extraction

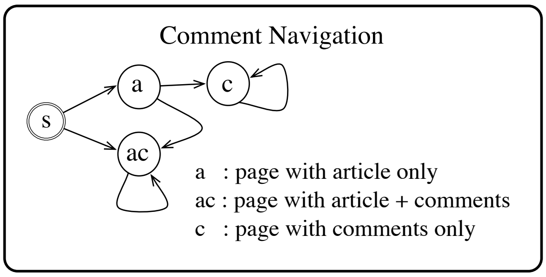
Comment Navigation
s
a
c
ac
a : page with article only
ac : page with article + comments
c : page with comments only

Figure 3: Navigational finite state machine

system. Similar to harvester, the extraction module employs an embedded browser. Although every site represents content differently, they all share navigational patterns: with at most one click from the landing page of an article, an initial set of comments can be reached; remaining comments can be then accessed either by pagination or expansion. See Figure 3. For example, `foxnews.com` requires the user to click on "Load more", and `huffingtonpost.com` provides pagination via "Next" links. All these actions, however, increase the processing time of the page: a `foxnews.com` article with thousands of comments might take an hour to load due to the wait time after each click action. Once a comment page is loaded in the browser data extraction takes place. For news content extraction, boilerpipe [7], a parser also available in Tika, efficiently and automatically detects the main article. However, for content presented in a list form such as comments or forum posts, *wrappers* provide better accuracy for our purpose as long as the page layout does not change. Our extraction module uses the *root* parameter, found in the *comments* section of the configuration file, to locate comment nodes. Then it iterates through the comment nodes by extracting the fields based on name-XPath pairs in *comment* section.

Aside from data such as author name and timestamp, some sites enable threaded discussion and present the dependency relationships with other comments on the page. Our data collection system preserves these conversational aspects by identifying the post to which a comment replies, e.g., reply-to (parent and child) relationships.

Our data collection effort faced a number of challenges.

- With some COI sites, a *requestRate* even greater than the default was needed to avoid detection and subsequent blocking of access.
- Many sites restrict discussion growth along either the time or depth axis, necessitating a fast turnaround between article detection and data extraction. For instance, `reuters.com` disables commenting after around three days; and `foxnews.com` refuses to even display comments after three days.
- Our XPath approach to extraction is sensitive to site re-design: during the course of our data collection we experienced one such incident.
- In order to do temporal analysis of the data, dates that the articles and comments are published need normalization. However, every site employs a different format and precision for date stamps: out of twenty sites we identified hundreds of distinct patterns for date formats.
- Especially for sites with threaded commenting, we had to perform extraction over the entire content to determine new comments during our revisits. Previously seen comments were identified by their unique ids within site. These ids are exposed in HTML source to support functionalities such as spam reporting. Content based duplicate detection failed due to changing user signatures.

Our current corpus was collected on hardware with 4x2.66GHz dual-core processors and 32GB RAM, running Linux. All software modules were developed using Python. We chose the document database MongoDB[5] as our storage. We reached to 100GB of database size, including indexes necessary to support crawling. For analysis purposes, we stored snapshots on a local machine and created more elaborate indexes. Firefox was used as an embedded browser and controlled via the Python selenium web driver[6].

4. LINGUISTIC ANALYSES

In the previous sections, we described the corpus and the manner of its collection. In the remainder of the paper, we report preliminary research concerning the linguistic and structural patterns associated with meme propogation in online communities.

4.1 Discussion structure

The comments associated with an article form a *discussion tree* with a structure that can vary in shape and size depending on various factors. Learning how and why a discussion grows is helpful in understanding the underlying community and the spread of ideas.

In sites that permit comment–comment attachment, we can observe the propensity of authors to reply directly to other authors versus

[5] http://www.mongodb.org

[6] http://seleniumhq.org

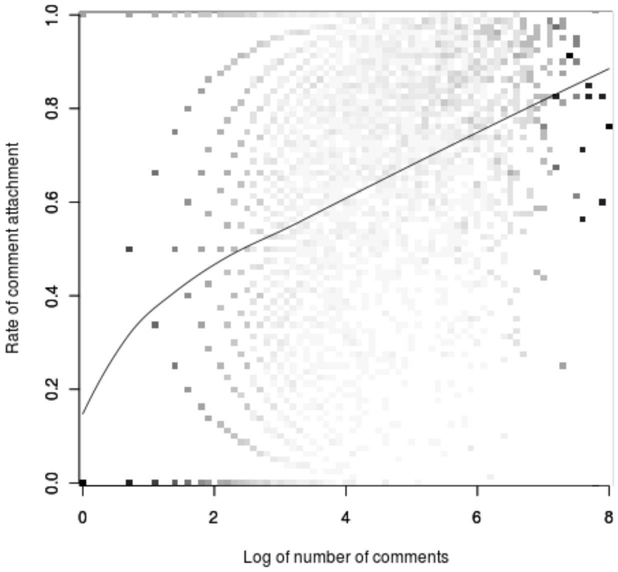

Figure 4: Rate of Comment Attachment, by Frequency of Posting, on foxnews.com

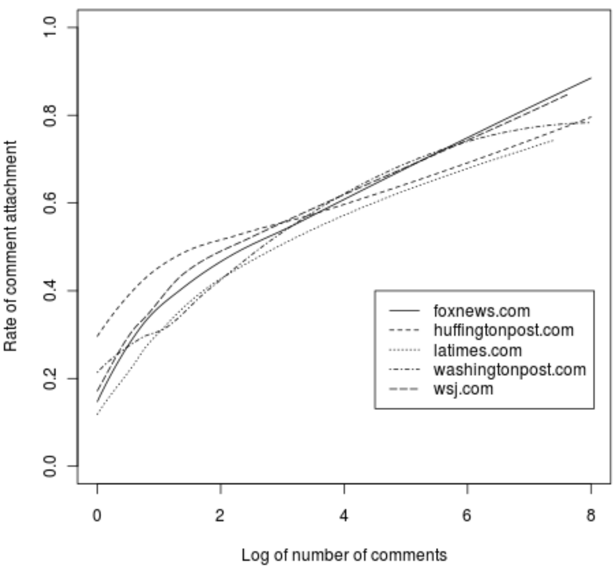

Figure 5: Smooth trendlines for Rate of Comment Attachment, by Frequency of Posting

replying to the root article. We have found a split of roughly sixty percent of comments attaching to other comments and forty percent attaching directly to the root node. Moreover, at the author level, we notice a clear relationship with frequency of posting: authors who post more frequently are more likely to attach to other comments as opposed to attaching to the root. Authors who have submitted exactly one comment to a website are much more likely to have replied to the root than to another author. Figure 4 plots the observed rate of attachment to other comments, by number of comments posted, for all authors posting to `foxnews.com`, along with a smoothed trendline. (The scatter plot is rendered in gray scale, with darker points representing higher conditional probability given number of comments.) Figure 5 demonstrates that this upward trend is characteristic of all sites studied.

What explains this behavior? It is likely that as an individual becomes more familiar with a website community, he or she is more willing to engage other discussants in debate. Individuals who post infrequently are more likely to have recently joined, and may therefore be uncomfortable participating in discussion. By contrast, the 'debater' community will inevitably contain individuals who are unwilling to let others have the last word. Another observation is that among the participants who never bother to reply to—or even read—others' opinions, few are likely to contribute more than one comment to any discussion.

4.2 Predicting Popularity and Attachment

Another of our goals in studying these data is to determine which words, phrases, or other linguistic forms influence the amount of attention the article receives, i.e., it's *popularity*. We describe here a simple experiment with this goal in mind. We formulate the experiment as a machine learning prediction problem involving a regression where the predictors (independent variables) are words in the article text, and the output (dependent variable) is the number of comments the article receives. We study 2,275 discussions from `wsj.com` occurring in July, August, and September of 2011, training the predictor on 1,820 (80%) of the discussions, and testing it on 455 (20%).

A central challenge to our prediction problem (and many others like it) is that the number of unique word forms in the data far exceeds the number of training samples. (The 2,275 articles in our current dataset contain approximately 60,000 unique word forms.) This means our focus needs to be on limiting the size of our model (regularization) in order to avoid overfitting. Support vector machines (SVMs) are the state-of-the-art in classification and regression in this setting. Interpreting their results, however, proves problematic due to a lack of an easily interpretable relationship between input features and model parameters. We therefore resort to sparse regression techniques and feature selection approaches, a method that has recently proven successful in the context of predicting demographics from social media [4]. Specifically, we focus our effort on an advance in this area called the elastic net [15], which is a regression technique that employs a regularizer that is a linear combination of the L_2 norm (ridge regression) and L_1 norm (lasso). We use the R package `glmnet` to fit the model and run our experiments.

We find that elasic-net regression works remarkably well in comparison with SVMs, outperforming it in terms of mean prediction error. After transforming the output variable (number of comments) into a quantile (thus normalizing the output variable to the interval $(0,1)$), the learned elastic-net regressor achieves a mean error of .176 on the training set. On the other hand, SVM regression[7] achieves a mean error of .264. Analysis of the learned sparse regression shows, for example, that for the selected period of `wsj.com`, the words *obama*, *taxes*, and *republicans* are the most

[7] We used the libsvm package with the nu-SVR option. The nu parameter was optimized on the test set by grid search.

effective positive predictors of popularity.

We have also begun to investigate the factors that drive the comment–by–comment growth of discussions, i.e., *attachment prediction*. Our goal here is to explain why particular branches of a discussion attract different amounts of attention. Viewing the growth of a discussion as a series of attachment decisions, we can ask which of the current nodes (including the original article, or root) is most likely to receive the next reply. We can formalize discussion formation as a generative process, in which each attachment decision is made according to an unknown distribution over existing nodes, and search for distributional models that best account for the observed sequence of attachments. A natural performance metric under these assumptions, borrowed from language modeling, is the attachment *perplexity*.

Our experimentation in this area involved comparison between three simple models. The first is a baseline uniform attachment model, which considers every comment (as well as the article itself) equally likely to receive the next comment. Second, we considered a *preferential attachment* model, which assigns each comment a probability that is proportional to the number of comments already attaching to it. This latter model was then refined by incorporating our prior findings about root attachment probability—we assigned a forty percent probability to root attachment, and split the remaining 60% probability among the remaining nodes according to preferential attachment.

The order in which these models are listed above corresponds to a consistent empirical ordering we observe on a range of datasets. Simple preferential attachment yields a considerably lower attachment perplexity than the uniform model, but is further improved by the model that recognizes the special status of root attachment. As we continue work in this area, we are searching for features of the comments themselves–their lexical content, say, or the identify of the commenter–that might allow us to refine further these simple models.

4.3 Community variation and contrast

Because our corpus draws from many different online communities, each with a large collection of authors, there is great potential in our corpus for studying linguistic variation amongst online communities. Additionally, the conversational (and often controversial) nature of the discussions provides a venue for studying contrasting ideologies amongst groups. This section describes some of our preliminary work in this area.

Exploratory analysis of lexical counts in comments shows that the community of commenters within each website in our corpus has remarkably distinct patterns of language use, particularly amongst the COI forum sites. Importantly, the observed distinctions go beyond thematic variation (e.g., differences in topic of discussion), and suggest marked *sociological* and *stylistic* contrasts among the sites. For example, we find a large variation in the frequency of personal pronouns, e.g., *you*, *I*, and *we*, which are known to be stable high-frequency indicators of genre [12]. For example, the word *I* ranges in frequency from 0.95% to 3.75% (a factor of 3.94) across all sites, and the word *we* ranges in frequency from 0.15% to 0.68% (a factor of 4.59).[8] We also measured vocabulary size by randomly sampling 100,000 words from the comments on each site and counting the number of unique words present in the sample (we report mean results from repeating this procedure 100 times). The resulting figure ranged from approximately 11,500 for the websites `motherjones.com`, `npr.org`, and `richarddawkins.net`, to below 9,000 for `mothering.com`. Both types of analysis suggest strong distinctions between sites.

Table 3: Psychometric analysis of websites using LIWC [14] word class counts.

LIWC class	Example	Site with greatest relative freq.
ANXIETY	*'worry'*	`mothering.com`
FRIENDS	*'buddy'*	`animalsuffering.com`
SWEARING	*'piss'*	`vanguardnewsnetwork.com`
CERTAINTY	*'always'*	`hindudharmaforums.com`
DEATH	*'bury'*	`animalrightsdiscussion.com`
INSIGHT	*'think'*	`richarddawkins.net`
NEG. EMOTION	*'ugly'*	`reuters.com`
POS. EMOTION	*'nice'*	`vegansoapbox.com`
INHIBITION	*'block'*	`conservativesforum.com`

By analyzing counts of psychologically-relevant words, the data also suggest distinct *psychological* characteristics of website communities. In particular, we use a dictionary of word classes distributed with the Linguistic Inquiry and Word Count (LIWC) software program [14]. LIWC is a system that performs psychometric analysis using counts of human-authored (and experimentally validated) word classes such as FAMILY, POSITIVE EMOTION, and CERTAINTY. Table 3 shows the results of applying a simple LIWC analysis as follows. For a collection of LIWC word classes, we list the website for which the word class has this highest relative frequency. We find these results to match our intuitions about community psychology that have been gained from direct experience with the comments.

4.3.1 Within-site contrasts

The analyses just described confirm that our corpus covers a wide variety of communities. However, one of the central hypotheses we ultimately wish to test with these data is that coherent but contrasting ideological communities exist *within* each site. For example, we expect that a controversial site like `richarddawkins.net` will contain many debates between Darwinians and creationists, and we want to be able to characterize the language use of these two ideological communities.

As a preliminary test of our hypothesis, we perform a simple analysis that contrasts the coocurrence of three-, two-, and one-word phrases at varying levels of discussion structure. The technique works by measuring how frequently two phrases co-occur in the same *discussion* and contrasting this with how *in*frequently they co-occur in the same *comment*. This allows us to identify pairs of phrases that play opposing roles within conversations. We measure this phenomenon using what we call the *bifurcation* of two phrases x and y such that

$$\mathrm{bifurcation}(x,y) = \mathrm{npmi}_{\mathrm{discussions}}(x,y) - \mathrm{npmi}_{\mathrm{comments}}(x,y)$$

where $\mathrm{npmi}_z(x,y)$ is the normalized pointwise mutual information of the occurence of phrases x and y in the collection of corpus units

[8] The word *I* occurs at least 10,000 times in each of our websites and the word *we* occurs at least 3,000 times.

specified by z such that

$$\mathrm{npmi}_z = \mathrm{pmi}_z(x,y)/-\log[\max(p_z(x),p_z(y))]$$
$$\mathrm{pmi}_z = p_z(x,y)/p_z(x)p_z(y)$$

where $p_z(w)$ is the proportion of units z in which the phrase w occurs, and $p_z(w_1,w_2)$ is the proportion of units z in which both phrases w_1 and w_2 occur.

The results of applying this analysis to our corpus show that some interesting word pairs can be found. From an analysis of `latimes.com`, for example, we find phrase pairs with high bifucation such as “tea party movement”—“tea party people.” This pair seems to reflect a positive and negative form of expression for the Tea Party.

It is apparent, however, that our approach warrants some refinement, and the bifurcation analysis also draws out some interesting but unexpected results. For example, the technique reveals phrase pairs like “end of story”—“matter of fact” and “thanks in advance”—“hope that helps,” both of which are indicative of discourse structure rather than opposing ideologies. Also, we find that the technique is very good at distinguishing foreign language comments. Analysis of `animalsuffering.com`, for example, revealed several discussions in which comments were in both English and French, generating word pairs like “she”—“elle.”

4.4 Linguistic meme epidemiology

Our discussion forums provide illustration of memetic outbreaks, in which an idea or attitude propagates throughout a website’s readership. The biological metaphor is apt: a community of susceptible individuals is exposed to an idea expressed by an “infected” individual. Some individuals are “immune” and do not spread the idea, whereas others readily adopt the meme in subsequent posts, becoming propagators of the idea. In discussion forums “exposure” occurs when one individual reads a posting in which an “infected” individual expresses the meme; the contact may or may not result in transmission. As more infected individuals express the meme, the number of contacts and hence infections increases and an “epidemic” ensues. The rate at which contact leads to transmission is dependent on the likelihood that an individual post is viewed by other community members, as well as the attractiveness of the idea being expressed. In many cases the epidemic subsides as infected individuals “recover”, no longer interested in active expression of the meme. The duration of the epidemic is affected by this recovery rate.

We have identified a number of memes that have attained currency during the time periods spanned by our collections. These include pithy epithets such as *Party of No* and catchphrases such as *once great nation*. Our investigations focus on linguistic memes: phrases or lexical entities that can be readily recognized in comments and transmitted with little loss. An example is the family of insult words containing the *-tard* suffix, such as *libtard* or *religiotard*. Starting in approximately 2007, when this phenomenon was virtually non-existent, the use of *-tard* as a general-purpose pejorative particle has seen rapid increase in several discussion forums. Our analysis pools all of these forms into a single lexical meme. Another example of a lexical meme is *cretinist*, a derogatory form of the word *creationist*.

Figure 6 depicts the number of authors expressing the *-tard* lexical meme on the `richarddawkins.net` site as a function of time. The adoption curve exhibits the classic shape of epidemic growth: rapid initial increase, peak, and gradual decay. Figure 7 illustrates the temporal growth in the number of authors expressing the idiom *kick the can down the road* on the `wsj.com` site. The difference in time scales for these two epidemics is noteworthy: evidently the *kick the can* phrase went out of fashion relatively quickly.

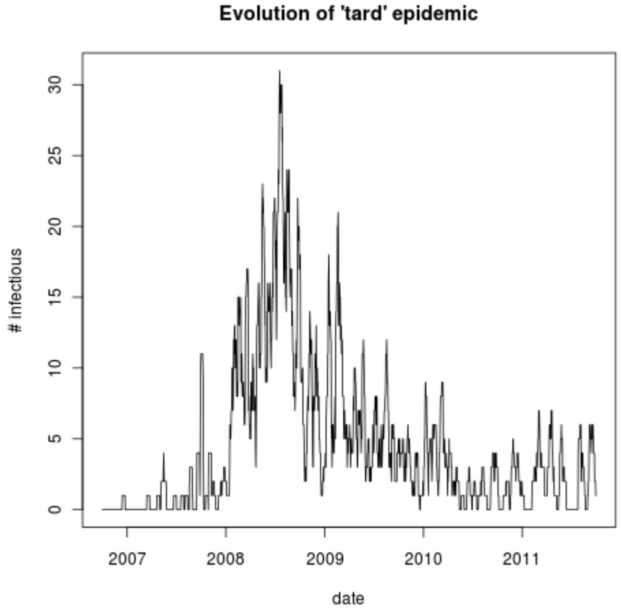

Figure 6: Adoption curve for *-tard* lexical meme

Statistical and epidemiological techniques can be applied to model meme outbreaks. Using a synchronic approach we might seek factors that predict some measure of severity of an outbreak, or that predict the chance that an individual will be receptive to a particular meme; a diachronic approach might predict the evolution of an outbreak as a function of history.

Our diachronic appproach adapts the familiar compartmental models from epidemiology to describe the dynamics of meme adoption. We have found that the classic SIR model yields a qualitatively compelling fit to the observed adoption curve for a number of meme outbreaks: If $x(t)$, $z(t)$, and $w(t)$ denote the number of susceptible, infectious, and recovered individuals at time t, then the growth of these populations is modeled by the following system of equations:

$$\begin{aligned}\dot{x}(t) &= -ax(t)z(t)\\ \dot{z}(t) &= bx(t)z(t) - dz(t)\\ \dot{w}(t) &= dz(t)\end{aligned}$$

for parameters a, b, and d. This classical epidemic behavior is often seen with novelty lexical memes, which tend to enjoy periods of popularity and subsequent decline that are largely unaffected by external events. For example, Figure 8 displays the fit of the SIR model to the *tard* epidemic observed at six-month intervals beginning in January 2008. Other memes are observed to follow the classic trajectory, or to exhibit a “steady-state” background rate of expression, but later experience a resurgence in popularity because of an external event (e.g., a news item) that heightens the visibility of the meme, and therefore alters the dynamics of adoption.

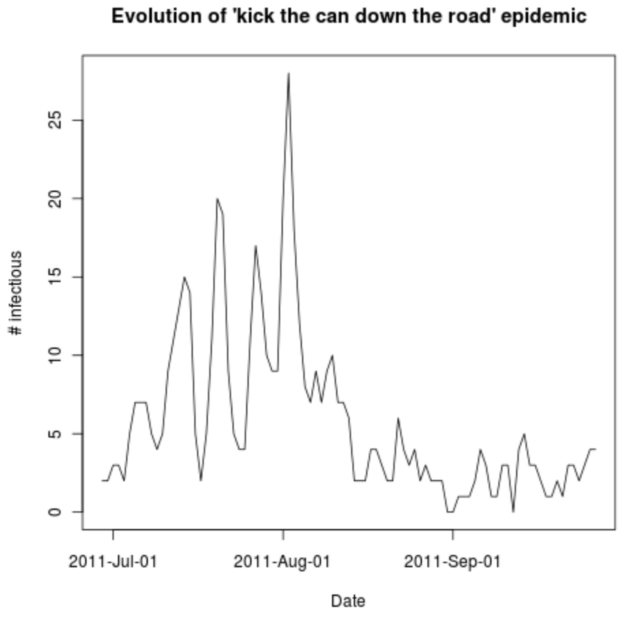

Figure 7: Adoption curve for *kick the can down the road*

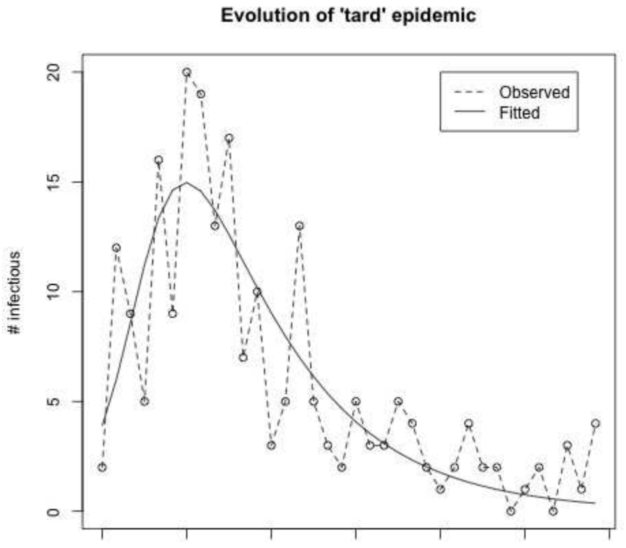

Figure 8: Observed and fitted adoption curve for *-tard* lexical meme

5. DISCUSSION AND FUTURE WORK

Online discussions provide fertile new ground for linguistic and socio-linguistic research, but the collection of such discussions poses more challenges than comparable social media. As this paper describes, we have worked through these challenges and amassed, over the course of a few months, a corpus of approximately half a billion words.

This data differs from other social media content in ways that open new avenues of investigation. Unlike blogs, there is little or no expectation that online comments will be carefully constructed or even grammatical. Unlike Twitter, discussants are not consciously broadcasting to the world, but are engaging in exchanges with the author of an article or other discussants. Unlike Facebook, discussants face no implicit pressure to maintain an identity. The low barrier to participation, compared to these other forms of social media, makes online discussion arguably more inclusive, contributing to a sample of linguistic utterance from a much broader demographic spectrum. Finally, the author-directed attachment of comments to an article or other comments gives rise to an interesting collaborative multi-document structure, the discussion, which other forms of social media do not provide.

We have only begun to exploit the opportunity this data provides, attempting to account for the spread of ideas, of memes, as an epidemiological phenomenon. None of this paper's sections offers the final word concerning its respective technical focus. Although our harvesting pipeline has assembled a corpus of considerable size, there are many lingering challenges, such as scaling to a larger number of sites, automating site acquisition and maintenance, and the normalization of comments to account for phenomena such as quoting or excerpting. We have by no means accounted for all the factors responsible for the rate and shape in which discussions grow. We surmise, for example, that different users have different effects on the propensity of a discussion to grow, some of it due to their language, and some to their identity. We see clear linguistic markers for community, but we have yet to measure the influence of community on the spread of ideas. And we have demonstrated the applicability of compartmental models to diachronic lexical adoption, but not more directly to the spread of ideas. All of these objectives remain the focus of future work.

6. ACKNOWLEDGMENTS

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory (AFRL) contract number FA8650-11-C-7119. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

7. REFERENCES

[1] L. A. Adamic and B. A. Huberman. Zipf's law and the Internet. *Glottometrics*, 3:143–150, 2002.

[2] E. Adar and L. Adamic. Tracking information epidemics in blogspace.

[3] C. Bauckhage. Insights into Internet memes. In *Proc. 5th Intl. AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 42–49, 2011.

[4] J. Eisenstein, N. A. Smith, and E. P. Xing. Discovering

sociolinguistic associations with structured sparsity. In *Proc. ACL 2011*, 2011.

[5] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010)*, 2010.

[6] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. *SIGKDD Explorations*, 6(2):43–52, 2004.

[7] C. Kohlschütter, P. Fankhauser, and W. Nejdl. Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 441–450, New York, NY, USA, 2010. ACM.

[8] P. L. Krapivsky and S. Redner. Organization of growing random networks. *Physical Review E*, 63, 066123, 2000.

[9] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proceedings of WWW 2003*, pages 568–576. ACM Press, 2003.

[10] K. Lerman and R. Ghosh. Information contagion: an empirical study of the spread of news on Digg and Twitter social networks. In *Proceedings of the Fourth International ICWSM Conference*, 2010.

[11] A. Mesbah, E. Bozdag, and A. van Deursen. Crawling ajax by inferring user interface state changes. In *Web Engineering, 2008. ICWE '08. Eighth International Conference on*, pages 122 –134, july 2008.

[12] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Text genre detection using common word frequencies. In *Proc. 18th Intl. Conf. on Computational Linguistics (COLING)*, pages 808–814, 2000.

[13] E. Sun, I. Rosenn, C. Marlow, and T. Lento. Gesundheit! Modeling contagion through Facebook news feed. In *Proceedings of the Third International ICWSM Conference*, 2009.

[14] Y. Tausczik and J. W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29, 2010.

[15] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67, Part 2:301–320, 2005.

Can Google count? Estimating search engine result consistency

Paul Rayson
School of Computing and Communications
Lancaster University
Lancaster, UK
+44 1524 510357
p.rayson@lancaster.ac.uk

Oliver Charles
School of Computing and Communications
Lancaster University
Lancaster, UK
oliver.g.charles@gmail.com

Ian Auty
School of Computing and Communications
Lancaster University
Lancaster, UK
i.auty@lancaster.ac.uk

ABSTRACT

In the last ten years, corpus and computational linguists have begun to source language samples from the Web. A standard pipeline has emerged for compilation of the 'web as corpus' including crawling, filtering, de-duplication, tokenising, indexing etc. However, there are certain areas where building a large enough corpus even from the web is not feasible, and it is tempting to use result counts derived from search engines to overcome the sparse data problem. In this paper we explore the stability of these search engine result counts for both multiword expressions and single words. Commercial search engines employ a range of techniques to estimate the counts, and thus it is important that researchers understand the implications and how to minimize this instability. Through a variety of different experiments and analysis, we investigate exactly how this stability manifests itself, and conclude with a set of guidelines on how future projects can ensure they are using accurate frequency data from search engines. Search engine result reliability will also have impact on corpora sourced from the web using the web as corpus paradigm.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *language models, text analysis.*

General Terms

Measurement, Reliability, Experimentation.

Keywords

Web as corpus, search engines, frequency.

1. INTRODUCTION

Interest in using the web as a corpus picked up momentum around 10 years ago [4] and at least four methods have emerged in which the Web can be used for language analysis [1]:

1. *The Web as a corpus surrogate* -- using the web as an interface to a corpus
2. *The web as a corpus shop* -- using commercial search engines to find corpora
3. *The web as corpus proper* -- using the entire web as a large corpus, at a point in time
4. *The mega-corpus/mini-web* -- using a subset of the web as a corpus

A lot of research has focused on these options – e.g. via the ongoing WaCKy project, which provides precompiled corpora from the web in various languages. Though the data is taken from the Web, it is only able to represent a snapshot in time, and search engine results play a significant part in determining the contents of the resulting corpora. In this way, language researchers can derive linguistic information from search engines by careful manipulation of the tools and APIs they provide. However, the agendas of commercial search engines and the needs of language researchers could not be further apart. Search engines are interested in returning 'useful' results to consumers, in the shortest time possible. This mismatch may be a significant problem and at least two aspects require further investigation. First, calculation of the exact total number of search results for a query is seldom carried out with the larger commercial search engines, rather an estimate is made based on a number of factors, and usually these determining factors are proprietary and not public knowledge. Secondly, the order in which results are returned may make a significant difference to the corpora collected from a web crawl.

Past research has looked at using search engines and their estimated result counts [2, 8] and shown mixed success. This previous research was all carried out near the start of the web-as-corpus boom, so re-investigating this topic is now appropriate - even five years is a lot of time for growth and change in the web. Google has continued its growth, Yahoo! underwent fairly radical changes, and Altavista lost market dominance (and is now owned by Yahoo!). A new competitor to the search engine market also came along - Bing, formally Live by Microsoft, and rapidly made itself known amongst the top three in terms of popularity.

In this paper, we investigate the issue of estimated result counts with a much larger data set than previous experiments, over a longer period of time, and derive a set of guidelines to help researchers use search engines reliably. In section 2, we explain in further detail the importance of search engine result counts, and how simply using the first retrieved number is rarely the safest option. In section 3, we describe a set of experiments that we carried out to determine the stability of search engines under various circumstances and periods of time, which we present and

- *WAC'7*, April 17, 2012, Lyon, France.

analyse in further detail in sections 4 and 5. From these results, we derive a set of guidelines in section 6 on how researchers can most responsibly make use of these estimated numbers. Finally, we conclude in section 7 with some final thoughts and potential future work.

2. BACKGROUND AND MOTIVATION

Search engine estimated result counts have many uses, predominantly in the natural language processing (NLP) domain. These frequencies are often used to aid in the cases whereby certain expressions cannot be found within a corpus. Keller and Lapata [3] researched this specifically in numerous publications, and are not alone in their work. Examples include machine translation between languages, spelling correction and adjective ordering [4]. Aside from the NLP domain, result counts are also useful in language learning, in order to help prioritize vocabulary for students.

While the WaCKy project does attempt to address these issues, it must be noted that due to the static nature of a precompiled corpus (vs. the web as a live corpus), it still does not truly utilise the dynamic nature of the Web. Furthermore, as the web is constantly expanding, it is obviously not possible for WaCKy to be fully representative. Hence, for some applications e.g. analysis of current news trends, online reputation management and internet advertising, it is useful to consider using live search engine result counts.

While there are clear domains where result counts are applicable, there is yet to be any concrete information on how to best attain search result counts, and be sure of the accuracy. We expect that search engines grossly estimate the result count for reasons of speed, and need to further investigate the wild fluctuations observed through informal observations whilst web searching.

3. METHODOLOGY

We have explored three different types of linguistic units: multiword expressions, single words and a small number of proper nouns to see if their search engine counts behave differently. Two sets of experiments have also been carried out over significant periods of time and nearly two years apart and this allows us to see if variability has changed over this period.

Our initial assumptions for the results were straightforward; none of the search engines would agree on result counts, and result counts would fluctuate over time. However, we hoped to see a fluctuation hovering around some sort of central value - a value that would most likely increase gradually over time as the indexed web expands. We also assumed that the approximations would trend in similar patterns, though we did not expect them to agree numerically.

The overall approach for the experiments was to regularly search for a large number of search terms over a fixed period of time. The estimated search result count, date, query and search provider would all be then logged, and finally analysed.

3.1 Collection and preparing data

For the first experiment, we took the top 1,000 single words from written and spoken English, from an analysis of the British National Corpus [6]. For the investigation of multiword expressions, we took a random sample of 2,000 expressions from an English semantic lexicon [7]. These multiword expressions varied in size, but the majority were 2-5 words in length. All the search engines we used provided the same syntax for entering queries so we encapsulated all queries in quotation marks, which specifies exact phrase matches (the words must be in the same order and distance apart).

For the second experiment, we wished to cast the net wider than high frequency words in order to see if similar variability emerged, hence we selected 10 high frequency words, and 10 low frequency words from the same frequency dictionary as above plus 10 names of towns and cities in the UK to provide some specific points of comparison.

3.2 Searching

We chose to search over three search engines: Google, Bing and Yahoo! At the time of writing, these search engines hold the majority of the market share in search engines [9], and all of these providers allowed for queries through a search API. By the time of the second experiment the API provided by Yahoo! was no longer free to use, so only Google and Bing were used at that point.

In the first experiment, a small Perl script was created to perform searches for all 3,000 queries every six hours, each day. The results were stored in an SQLite database, with a row for each query on each search provider, time-stamped with the time of searching. This process was carried out for approximately 20 weeks, through October, 2009 to March, 2010.

In the second more focussed experiment with a smaller number of words, we polled the search engines every 12 hours for a period of 5 weeks in November and December 2011. For this iteration, a Java program and a MySQL database was employed to record the data. In both experiments, timings for making use of the search engine APIs were chosen to avoid hitting limits for multiple requests.

3.3 The effect of regional searches

Alongside the searches above, we attempted to re-run the first experiment over search engines specialised for certain countries. We were interested to see if the stability of search terms was less (or perhaps) more influenced by limiting the results that would be obtained.

In order to do this, we ran the same queries over Google Australia, Ireland, Isle of Man, New Zealand and United Kingdom. We only investigated Google due to its clean API for these types of searches, but also because we were more interested in the consistency within a single search engine.

This regional experiment was run on a smaller scale, running throughout February 2010 to the end of March 2010. The searches we again carried out over six hour intervals.

3.4 Web interfaces

We initially approached this research with the assumption that a web interface does not use the same algorithms for estimating result counts as a dedicated API. We believed that the estimated search count on the web pages would be higher, or more grossly estimated than that of the API, as this number also serves as a marketing tool – people think higher is better.

To investigate this further, we took a sample of 30 multiword expressions from our first 2,000, and 30 single words. For each of these we performed the identical search by hand through a web browser. The results were then logged, and we compared these against the actual reported counts from the search API.

4. RESULTS FROM EXPERIMENT 1

One of the main forms of analysis we used was the shift width between two search terms. The shift width is defined as:

$$s_i = \frac{f_i - f_{i-1}}{f_{i-1}}$$

i was the number of days since the start of the experiment, and each shift was calculated between consecutive days. s_i is the calculated shift width; f_i is the current result count and f_{i-1} is the previous days result count. As these functions operate on the shift per day, we aggregated multiple searches per day (as searches were performed over 6 hour intervals) by taking the mean shift.

Variants of this formula were used to calculate shift width over all search terms of a search engine, or for a set of queries. For example, to calculate the shift over all single word queries, or for all queries on a single provider.

Originally, we did attempt to analyse this data through other techniques. Our first attempt was to display a sliding window of the standard deviation of the search result over a day, a week, two weeks and the duration of the experiment. The window size here was a critical factor, and the interesting data was only uncovered for the smaller window sizes - and was better displayed using shift widths. We also looked at displaying the total frequency over time - while this was not useful enough for our primary form of analysis (due to the large differences in scale), it has still been used to support the shift width data in some cases.

4.1 Trends within three search engines

To investigate trends within the three investigated search engines, we studied the interquartile range and median for the aggregated shift width i.e. the shift width for all terms. For each query used, the shift width was calculated over the course of the experiment, and then these results were averaged in order to find these statistics. Figure 1 shows the results of these calculations, over time for each of the three search engines.

Google shows itself to be almost constantly stable. The first two months of the experiment saw the interquartile range staying constantly between ±2%. The most striking result from figure 1 is the massive instability around the 6th - 14th January. Here we saw a shift up by 15% and then a drop by 8% - clearly illustrating the unpredictable nature of search engines.

Yahoo!, again, shows the unpredictable nature of search result counts, though at a different point in time. At the start of November, there is a large period of instability. Over the course of three days, Yahoo! shifted by 0.24 (a 24% increase) followed afterwards with a shift down by -0.14 (a 14% decrease). Due to no apparent correlation between these two periods of instability, we can only assume they were due to temporary system difficulties.

Despite the instability during the start of November, Yahoo! shows itself to be even more stable than Google, with the interquartile range consistently staying within ±0.5%.

Another interesting fact that can be seen within this graph is that there are large regions where the shift width appears to be below 0, which would signify an overall trend for the decrease of search results overtime within this period. This is analysed further towards the end of this section (see table 1 and related discussion).

Finally, Bing shows more contrasting results to Yahoo! and Google. Bing showed a lot of instability during the first two months of the experiment, though it has drastically stabilised since this time. Bing is the youngest of the three search engines, launched May 2009, while Yahoo! and Google were launched in 1995 and 1997 respectively. Due to this age difference, Bing may have had radically different indexes to Google and Yahoo!

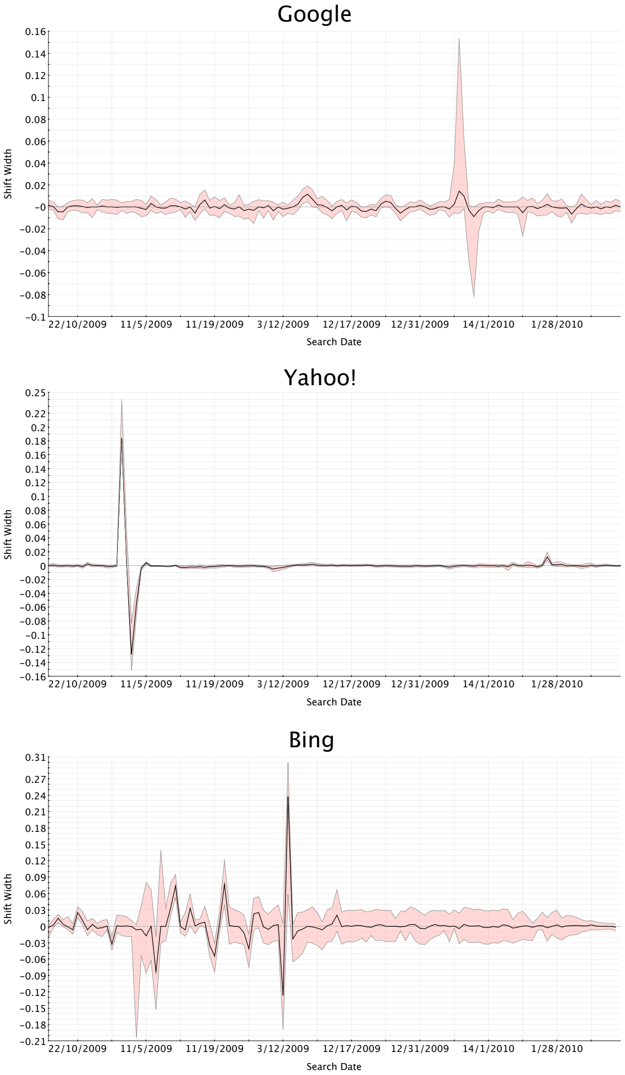

Figure 1: The median shift within the interquartile range for Google, Yahoo! and Bing (for all search queries)

To complement the results of aggregating shift widths together per search engine, we also investigated the growth (or contraction) of search results for each engine. Least square regression was used to find the regression of the search result count over time, and this was calculated for each query. We then aggregated these results and found the mean and standard deviation.

Table 1: Mean and standard deviation for the gradient of regression lines over search engines.

Provider	Mean	Standard Deviation
Google	-1121	13669
Yahoo!	8683	34307
Bing	68818	458873

As one would expect, these numbers are not particularly accurate to summarise either the change in search engines databases, or to measure the size of the Web, due to the large standard deviation. However, it does serve to illustrate the differences between providers – table 1 shows largely contrasting results for each provider.

Bing is especially suspect, with an estimated growth that is significantly higher than both Yahoo! and Google. One plausible conjecture is that Bing's infancy to the search engine market at the time of the first experiment required it to grossly estimate reported frequencies in order to account for the Web that they have yet to crawl. This conjecture is somewhat verified when one looks in detail at the spiking in Bing's shift width graph. We see many positive shifts, though not many negative shifts - which would indicate an increase in index size at this point in time.

4.2 Interesting queries

To discover interesting queries, a script was run to calculate the interquartile range of the shift width for each query, for each provider. With this list of interquartile ranges we were then able to select queries with a proportionally higher or lower interquartile range than other queries. We thus discovered a few interesting queries that highlight some of the potential problems with using search engine result counts.

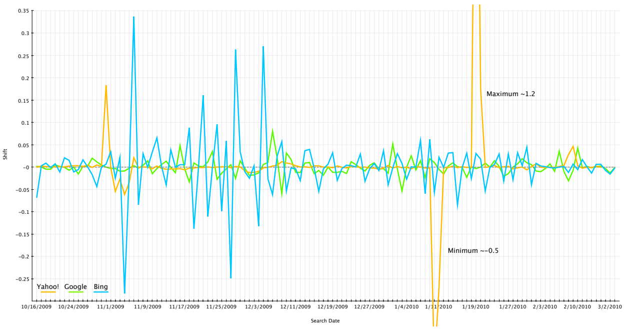

Figure 2: Analysis of the ``Best of both worlds'' query

The query "best of both worlds" was recorded as being fairly stable within Google, and thus we begun to explore this query in more detail (figure 2). The Google shift width for this query is in green, and consistently stays within ±0.05, indicating that search results do not change by much more than 5% between days. The Yahoo! curve itself (shown in orange) was even more stable showing almost no variance whatsoever - though there are some significant blips at the start of November and towards the end of January. Bing showed itself to fluctuate wildly between days, varying 35% between days for October and November, and stabilizing somewhat mid December (though still varying more than Google).

The key point to take away from this graph, is that search engines do not commonly correlate as far as stability is concerned. The minimums and maximums of these curves do not consistently match, and the larger peaks (for example Yahoo! in January) cannot be observed for other search providers.

The discrepancy between providers is emphasised further if we consider the search result frequencies themselves. Figure 3 shows the scales to be radically different between providers, with Bing estimating approximately 55 million results, and Yahoo! and Google reporting more conservative 320,000 and 1.6 million estimates respectively.

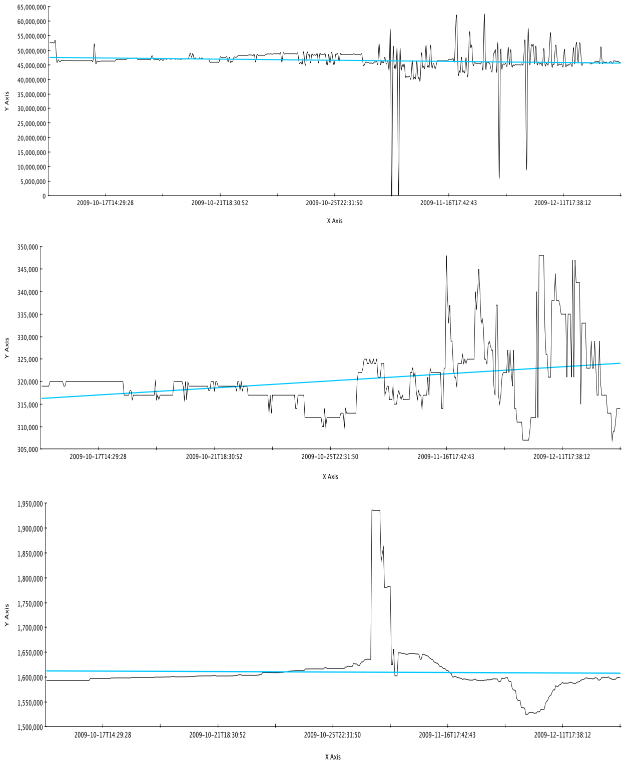

Figure 3: The "best of both worlds" query frequencies for Bing, Google and Yahoo! respectively

Figure 3 also captures something that cannot be seen within figure 2 easily, which is the gradual shift. Yahoo! in the end of December sees a shallow local minimum over the course of a few weeks - estimates taken during this period would not be very representative of the search term compared to a sample taken during October or early November.

Growth over the search engines was also not consistent. Bing and Yahoo! agree that the query is returning fewer results over time (though at different rates), while Google reports it as being fairly stable, neither increasing or decreasing in frequency.

As well as finding a multiple word query, we have also chosen a single word query for more detailed analysis. The query "bed" was shown to have an interquartile range for shift of approximately 0.3 for Google - which puts it outside the expected interquartile range of 0.0480 by a significant amount. Again, we see no correlation between the search providers - Yahoo! exhibiting a relatively flat and stable result count around 48 million, while Google shows more instability and also jumping from 18 million to 24 million, and then later back down. Bing estimates the result count an order of magnitude higher than both Yahoo! and Google, trending to around 350 million.

The shifts are clearer in figure 4, where Yahoo! sticks mostly to a consistent straight line as apparent in figure 5. However, this time Google and Bing both show signs of instability, with shift peaks moving past 0.05 fairly regularly. Bing shows more instability that persisted throughout the experiment. However, it would clearly be difficult to make use of the result count from Google as it does not seem to be stable for any usable length of time (with the longest stretch being a few weeks).

Alongside this detailed analysis, we also attempted to summarise the stability of all queries - separated as single-word queries and phrases. To achieve this, we calculated the interquartile range for each query, and then aggregated these results to find the mean and standard deviation. Table 2 shows the results of this analysis.

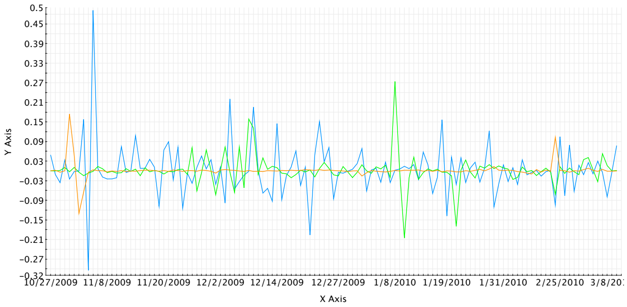

Figure 4: The shift for the query "bed" over all three search providers

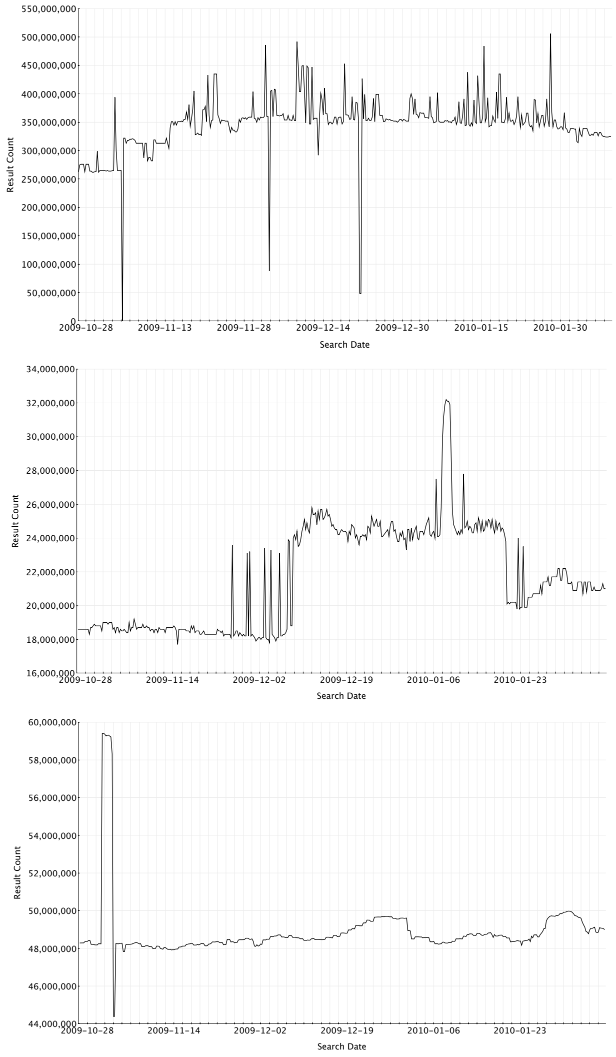

Figure 5: The "bed" query frequencies for Bing, Google and Yahoo! respectively

Table 2: A summary of the interquartile range for both single word queries and multiple word phrases

Provider	Single words		Phrases	
	Mean	S.D.	Mean	S.D.
Google	0.0480	0.0283	0.1910	0.325
Yahoo!	0.0377	0.0291	0.0833	0.146
Bing	0.1330	0.2450	0.0728	0.194

Interestingly, there is little correlation between the 3 search providers. For both Google and Yahoo, a single word query was on average more stable than a multiple word query, exhibiting a lower mean shift interquartile range, signifying that shift generally stays under 5%. However, Bing actually showed more variation in a single word query than a multiple word query - in direct contrast to Google and Yahoo!

It is promising to see that these shifts are low across all providers, and on average the shift stays within the same order of magnitude. Out of all multiword queries, we observed that for Google 94% of searches had a shift interquartile range < 1, while for Yahoo! and Bing 99% of multiple word queries had a shift interquartile range < 1. For single word queries, we found even higher levels of stability, with 100% of all single word queries having a shift interquartile range <1 in Google and Yahoo! and 97.4% in Bing.

4.3 The effect of regional-specific searches

After running the regional searches for a few weeks, we did not observe any noticeably anomalies between the search engines – as shown in figure 6 they all consistently trended in the same direction. We saw some minor separation towards the end of March, though this was less than 1% even between the biggest difference (Google Australia and Google UK).

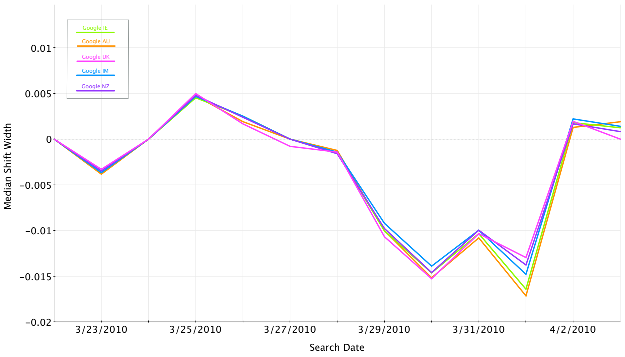

Figure 6: The median and interquartile range shift width over regional Google searches. From top to bottom, left to right: Google UK, New Zealand, Isle of Man, Ireland, Australia and worldwide.

From this we can conclude that, at least within Google, the regions are all indexed in almost identical ways. Providing the primary language is the same, we did not see any major differences in how numbers changed.

Interestingly, we can also see the close similarity when we look at the search result counts observed for each regional Google version. Table 3 shows the mean result count for each region, over all queries. While we do see variation, all means were in the same order of magnitude - a finding that is particularly suspect when we consider the contrast in population sizes for these countries. It is clear that the results are not country specific, otherwise it would mean for example, the Isle of Man has an estimated population of approximately 80,000, while Australia is estimated to have over 22 million.

Table 3: The mean result count for all search providers over the course of the experiment

Region	Mean result count
Australia	32,342,500
Ireland	29,592,290
Isle of Man	28,555,398
New Zealand	29,135,804
United Kingdom	28,330,142

From table 3, we can see that Google does not separate these regions such that it results in drastically different population counts. It is hard to understand exactly what is happening, as almost all of this information is proprietary to Google, however we can conjecture that as the results are in the same language they are most likely stored on the same cluster of servers, and as such the data is shared to make estimated result counts. The difference with regional searches is in the presentation of the results, but we have not found proof that it affects the estimated results count.

4.4 The differences between front-ends and APIs

For the majority of people, their primary mode of interaction with a search engine is via the site using their web browsers. However, this is not the only way to search for queries - it is also possible to use the APIs which provide a programmatic interface to the search engines which is convenient for developing software. We initially chose to use a search engine API due to terms and conditions, but as we began to analyse the gathered data it seemed different from that which was reported on the website front-ends.

Table 4: The relative changes between searching with an API and a Web front-end

Provider	Web front-end multiplier	
	Single words	Phrases
Google	9.86	9.71
Yahoo!	19.08	17.59
Bing	2.24	0.32

What we found, as presented in table 4, confirmed our expectations - Google and Yahoo! both appear to grossly over estimate the displayed search count, when compared to the API result. However, we do not see the same with Bing - which would indicate that the API and front-end share very similar (if not identical) algorithms in order to estimate the result count.

Comparing searches for multiple words compared to single word queries shows a small, but consistent discrepancy. A search for multiple words yielded a smaller difference from the frontend (though there was still a difference in all cases). A search for single words showed a larger difference, with the most pronounced change in Yahoo! which estimated 20 times more results on the Web front-end than in the API.

5. RESULTS FROM EXPERIMENT 2

Having considered large numbers of words and gross shift patterns in the first experiment, we wished to focus the second experiment on a smaller number of words in order to facilitate examination of relative trend patterns between words. As stated above, we selected 10 high frequency and 10 low frequency words plus 10 names of towns and cities in the UK. Also, by the time of the second experiment, November – December 2011, the Yahoo! API was no longer freely available so we observed estimated search result counts from Google and Bing only.

Figure 7 shows a snapshot of the estimated search result counts in Bing and Google of the 10 high frequency words. The overall patterns are generally similar although it can be seen that 'by' and 'that' have different positions relative to the other words. In addition, the frequency estimates for Bing are roughly half that of Google for the same words.

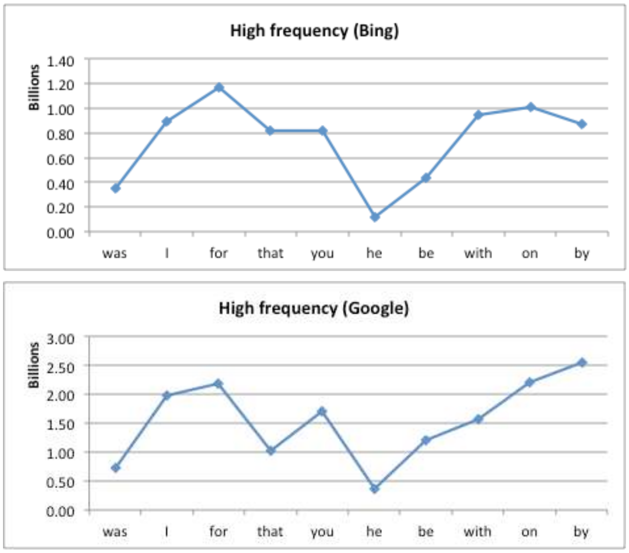

Figure 7: 10 high frequency words sampled on 18th December

In contrast, the 10 low frequency words (figure 8) show more difference in their relative patterns between Bing and Google, with 'boost', 'exchanged' and 'graph' showing opposite trends. There is a difference between the frequency estimates once again, however this time Bing's estimates are 10 times higher than Google's for the same words.

It is perhaps not surprising that estimates for lower frequency words are more variable than high frequency words since occurrences for generally less frequent words are more dependent on the text that has been indexed. However, taken together these two sets of results show conflicting trends and provide a strong hint that search result count estimates are calculated by very different mechanisms in Google and Bing.

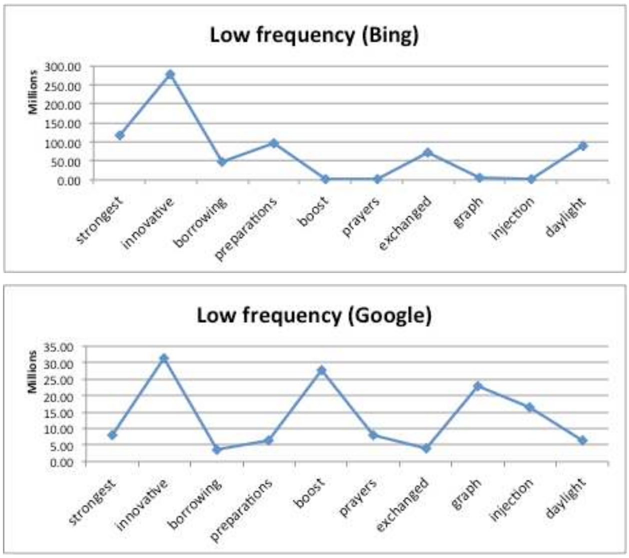

Figure 8: Low frequency words sampled on 18th December

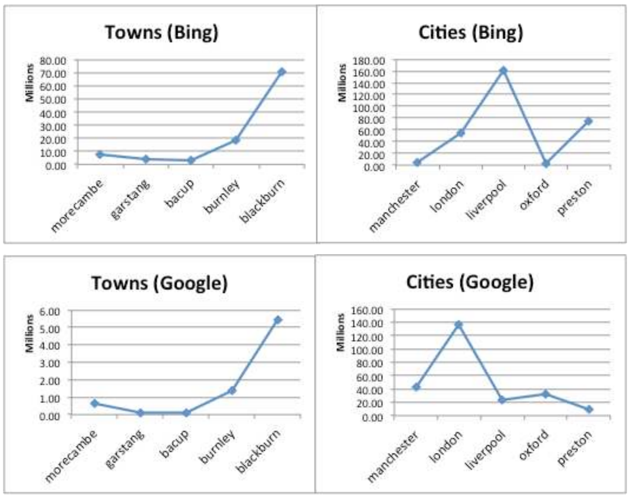

Figure 9: Estimated search result counts for 10 towns and cities

Figure 9 shows snapshots of estimated counts for 10 towns and cities. Bing's estimated results for towns are around 10 times higher than Google's although they show the same trends. In contrast, the pattern observed for 5 cities is radically different. Bing surprisingly estimates more search results for Preston and Liverpool compared to London whereas Google's results are more in line with expectations. Bing's results could possibly be a result of personalisation by geographical location.

6. GUIDELINES

In this section we summarise a set of guidelines that researchers can use in order to reliably use search result counts in the future. These guidelines have all been motivated directly from the results we observed in sections 4 and 5.

6.1 Do not trust web interfaces

Results from the consumer front-end of search engines are often highly misleading, inflating results drastically when contrasted against the results from the API. Comparing the search queries from our collected results, we observed estimated results counts on the first page of results to be estimated in the millions - drastically contradicting the results gathered from querying the API, as verified in section 4.3.

Careful exploration of the results from the front-end, however, shows that the first page is indeed an overly optimistic estimate. As one continues the search over multiple pages, the estimated count drops, until finally agreeing with the result from the API. However, this is not true for Bing and perhaps other newer search engine providers - this was only found true for Google and Yahoo! We recommend that researchers carry out some preliminary tests if they are going to use another search engine to see how the API behaves in comparison to a Web interface.

We suggest avoiding this discrepancy entirely, and consulting the API only. However, if you must use a search front end (for example, the search provider does not provide an API), you must consider more than just the first page of results, and be sure to consult the last pages of the results as well.

6.2 Consider a variety of queries

It is important to be aware of the two reasons for discrepancy within search results - fluctuations at per-query, and fluctuations within some of the search engines servers. In order to be able to distinguish between these two, we encourage researchers to test with an ample amount of queries (we recommend at least 15) and compare them. Do they all shift at the same time? If so, it may well be an unexpected change on the server side, and you should consider collecting more data.

6.3 Be aware of other search providers

Each search engine is radically different, as was shown looking at queries in sections 4.2 and 5. We encourage researchers to consider carrying their searches out with at least two search engines and compare and contrast the results. If the search results show similar shifts - either shifting together or shifting by relatively small amounts then the results may be used with confidence that this is as accurate as an estimated result is likely to be. However, if a provider is reporting unstable results, it may certainly be worth prolonging the searches for further investigation.

6.4 Observe search results for at least two weeks

In section 4, we have seen instabilities over various small lengths of time. Taking a single result, there is no indication whether the estimate is representative of the search provider's data. However, we did notice anomalous data that continued to be present for more than a few days. As such, we suggest that a two week window is used for searches - attempting to search every day. By calculating shift widths or using other statistical techniques, it is now trivial to identify the expected estimated result count for the query, and those estimates which are outliers.

6.5 Attempt to compare to other established corpora

In sections 4.3 and 5 we saw that the result count reported by search engines was drastically different for the queries we investigated. As such, we recommend attempting to calculate scaling parameters by comparing search result counts from a search engine to the counts in a corpus if possible. One would expect that over time this scaling will change, so it would be meaningless to provide a value now. Researchers should attempt to calculate this value as soon as possible to beginning their experiment.

7. CONCLUSION

Studying such a vast amount of data, within three different search engines, and over a long period of time has proved an illuminating experience. We have begun to make a strong move forward into understanding how these counts change. Furthermore, our guidelines do show circumstances whereby researchers could make use of search result counts and how they can ensure the accuracy of this data.

Eu (2008) attempted similar research on a smaller scale, and the results here are consistent with the earlier study but on a larger scale. We noticed Yahoo! tended to show less shift over time, and the frequency of very large shifts was infrequent. Eu argues that it is risky to conclude much about the stability of search engine result counts, however since then it is promising to see that the situation is at least consistent, if not improving, with that of previous years.

In our studies we sampled search engine counts over a long period of time, though we used some relatively straightforward statistical descriptions. If more detail was required, time series analysis methods could provide yet more insight. The use of time series analysis may help discover underlying patterns within the search result counts.

A useful future experiment could ascertain the real-world practicality of using estimated result counts by performing an application-based evaluation against frequency information from pre-existing corpora. A simple example could be a spell checking application where one could compare the accuracy of a spell checker that infers word frequencies from search engines to one that determines them from a known corpus.

We investigated search engines by three different query sets – single words, multiword expression and selected proper nouns, however researchers may be interested in yet more rigorous analysis. Future work could focus on observing relative result count stability i.e. do ranks in a resulting word frequency list change over time.

Search engine result instability also has possible implications for corpus data collected from crawls. Resulting corpora may not be replicable across two search engines or between two time periods. Therefore, the results described in this paper will have wider impact on those collecting data via search engine results as well as those using result counts as proxies for frequency estimates.

8. REFERENCES

[1] Silvia Bernardini, Marco Baroni, and Stefan Evert. A WaCky Introduction. WaCky, 2006.

[2] Jinseung Eu. Testing search engine frequencies: Patterns of Inconsistency. *Corpus Linguistics and Linguistic Theory*, 4(2):177–207, 2008.

[3] Frank Keller and Mirella Lapata. Using the Web to Obtain Frequencies for Unseen Bigrams. *Computational Linguistics*, 29(3):459–484, September 2003.

[4] Adam Kilgarriff and Gregory Grefenstette. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29(3):333–347, September 2003.

[5] Mirella Lapata and Frank Keller. The Web as a Baseline: Evaluating the Performance of Unsupervised Web-based Models for a Range of NLP Tasks. In *HLT/NAACL proceedings*, pages 121–128, 2004.

[6] Geoffrey Leech, Paul Rayson, and Andrew Wilson. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Longman, 2001.

[7] Scott S.L. Piao, Dawn Archer, Olga Mudraya, Paul Rayson, Roger Garside, Tony McEnery, and Andrew Wilson. A Large Semantic Lexicon for Corpus Annotation. In *Corpus Linguistics Conference 2005 proceedings*, 2005.

[8] Ronald Rousseau. Daily Time Series of Common Single Word Searches in AltaVista and NorthernLight. *Cybermetrics*, 2/3(1), 1999.

[9] SEO Consultants Directory. Top ten search engines. http://www.seoconsultants.com/search-engines/.

Using Web Corpora for the Recognition of Regional Variation in Standard German Collocations

Tobias Roth
University of Basel
Deutsches Seminar
Nadelberg 4
Basel, Switzerland
tobias.roth@unibas.ch

ABSTRACT

The Standard German collocation *berührende Worte* 'touching words' is only used in Austria. Yet, it consists of exclusively Common German component words. It is the combination that makes it a regional variant — one on a purely collocational level.

One of the goals in our German collocations dictionary project is to describe regional variation in the collocations collected. The main tool for this is a web corpus divided into three subcorpora. Each of them contains web pages from a different country code top-level domain (Austria, Germany and Switzerland).

Our method of variation recognition using a web corpus is compared to previously used resources. Advantages, disadvantages and results of the web-corpus approach are discussed. Given that the extent of regional variation in German collocations and its structuring into collocational, lexical and syntactical variation has been unknown, we present first estimates in that respect.

Categories and Subject Descriptors

J.5 [**Arts and Humanities**]: Linguistics—*corpus linguistics, German language, regional variation, collocations*

General Terms

Algorithms, Experimentation, Verification

1. INTRODUCTION

Regional variation within Standard German is not limited to the lexicon, to pronunciation, to orthography or syntax to name the most prominent fields where variation has been inquired. It is also relevant when dealing with collocations. Collocations do not inherit all of their regional variation from their lexical components or their syntactical structure. There is also regional variation on a collocational level.

Our very practical reason for dealing with the topic — and our main focus as well — is a lexicographical project in which we collect German collocations[1] and want to mark regional variants. We have decided to build and use a web corpus for this purpose.

In this article, first the concept of regional variation in Standard German is introduced as well as various types of variation. We will then present our web corpus and our collocation-extraction method. The next point will be how to recognise variation in collocations. That method together with our corpus is subsequently cross-compared with already known variation data. In the last section, we present own results, i. e. discuss the usage of the web corpus in our dictionary project and give quantitative estimates of regional variation in collocations.

2. REGIONAL VARIATION IN STANDARD GERMAN

Standard German (as opposed to German dialects) is a language with considerable regional and national variation. It is often referred to as a pluricentric language with different national and regional varieties [3]. According to [3], so-called *full centres* ('Vollzentren') are areas where German is an official language and shows internal codification. The German language currently has three full centres: Germany, Austria and Switzerland. They are distinct from *semi-centres* ('Halbzentren') where such internal codification cannot or not clearly be observed. Semi-centres of German are Liechtenstein, Luxemburg, South Tyrol and East Belgium. Variants from Germany (DE), Austria (AT) and Switzerland (CH) are often called *Teutonisms*, *Austriacisms* and *Helvetisms*, respectively. The term *Common German* denotes Standard German forms that are used throughout the whole language area.

There are a number of salient regional variants that speakers are conscious of, mainly on a lexical level. A bicycle in Swiss Standard German is a *Velo*, but the Common German *Fahrrad* can be used interchangeably. Or *Blumenkohl* ('cauliflower') in Switzerland and Germany is *Karfiol* in Austria. An important reference for German variants on a lexical level is the *Variantenwörterbuch des Deutschen* [4]. It also contains phraseologisms, but only very few collocations.

[1] See `http://www.kollokationenwoerterbuch.ch`.

2.1 Types of Variation

Different types of regional variation can be distinguished. Some of them are rather general and others apply specifically to collocations.

We have already established the lexical level as an example where variation can occur. However, variation in German is not limited to the lexical level. It comprises all levels of language. We can observe phonetic variation ("accents"), orthographic variation (e. g. the substitution of *ß* by *ss* in Switzerland), morphological variation (e. g. different diminutive suffixes: *-chen* in the North vs. *-lein* in the South) and syntactic variation (e. g. different auxiliaries for perfect forms with certain verbs: *sitzen* 'to sit' with *haben* 'to have' in the North vs. *sein* 'to be' in the South).

Variation on a collocational level is closely related to variation on a lexical level. As already mentioned, the lexical level is also the most prominent level of variation. Furthermore, there are different types of variation on this lexical level. One type of variant is the one where things and facts referred to only exist in a certain area or country. Therefore, the word is only used in that area, too. Typical examples are terms in connection with public authorities: *Zweisprachigkeitsprüfung* is only used in South Tyrol because it denotes a special language test that candidates for certain jobs in South Tyrol have to pass in order to prove their sufficient knowledge of two official languages. Or the parliament is called *Bundestag* in Germany only, not in other parts of the German-speaking area.

Another type of lexical variation and maybe the most prototypical one is the type in which there are different variants in different areas, all for the same concept. The variants may be mutually exclusive (such as *Alm* AT, DE and *Alp* CH 'alpine pasture'), but are not bound to be (see the bicycle example above with Common German *Fahrrad* and *Velo* in Switzerland).

Rather frequent is the case in which a word is Common German, but has an additional meaning in a certain area. E. g. the word *Haushalt* 'household' is Common German whereas *Haushalt* 'budget' is mainly used in Germany.

Leaving aside the structural levels for a moment, we turn to a another important perspective of variation: regional granularity. Variation exists in very different dimensions. In some cases, variants are used very locally only, in other cases they are used in several countries.

And a last important point is the fact that variants differ in absoluteness or frequency. Very often, a specific variant is not the only means of expressing a certain meaning in a certain area. It is often rather a matter of frequency and of higher or lower probabilities. A certain form may be found everywhere, but particularly frequently in one area, and thus constitute a variant of frequency.

2.2 Variation in Collocations

Different types of variation in collocations derive from the types of variation just discussed. Collocational variation can be inherited from lexical variation: because *Velo* is a Helvetism the collocation *Velo fahren* (bicycle drive, 'to ride a bicycle') necessarily is a Helvetism, too, even if the only difference from Common German *Fahrrad fahren* is the lexical variant for *bicycle*.

Another type of inherited variation in collocations comes from syntactic or morphosyntactic variation. If *Weihnachten* ('Christmas') is neuter singular in certain regions of Germany instead of the Common German plural, then it is not surprising that the formula *Merry Christmas* will be *Frohes Weihnachten* (with a neuter singular ending on the adjective) instead of *Frohe Weihnachten*.

The most genuine kind of variation on a collocational level is when all parts of the collocation are Common German, but the collocation itself is only used in a certain area. The collocation *tiefer Preis* (deep price, 'low price') — a Helvetism — is cited in [14] as an example for this type.[2]

We are particularly interested in this last type of variation, but for the lexicographical work, the other types will have to be considered just as well. In the collocations dictionary, we will give rather coarse-grained variation information, i. e. just AT, CH, DE for national variants of the three full centres, and no graded frequency information (so, no *esp. AT* or the like), as regional variation is just one aspect of the dictionary and not its main focus. This is also the reason why in this paper, we do not distinguish between regional and national variation, but take national variation as a subset of regional variation not explicitly mentioned but primarily aimed at in the majority of the cases.

3. CORPUS

3.1 Existing Corpora

There already exist several corpora of Standard German sensitive to questions of regional variation. Some of these corpora — such as the *Swiss Text Corpus* [8] and the *Korpus Südtirol* [1] — contain only text coming from a specific country or region. Others are not regionally limited but provide information about geographical origins of their texts, or this information is at least easily derivable. The largest German corpus *DeReKo* [17] enables users to explicitly choose their sources and even offers a view where results are displayed by national origin. Another corpus particularly suitable to investigate regional variation in Standard German is *Korpus C4* [10]. It is a distributed corpus with texts from Germany, Switzerland, Austria and South Tirol (Italy), reaching about 50 million running words[3]. It consists of parts of *DWDS (Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts)* [13], the complete collections of *Swiss Text Corpus* and *Korpus Südtirol* and a small part of the *AAC (Austrian Academy Corpus)* [7].

3.2 Yet Another Corpus

In spite of these already existing resources, we decided to create our own corpus. The two main reasons for this decision are related to size and availability of the existing corpora. For the study of collocations you need corpora of

[2] This type is not always easy to distinguish from cases where there is just variation in the meaning of one collocation member.

[3] "Running words" defined as space-separated tokens excluding punctuation marks.

considerable size so that you have enough text to be able to apply statistical association measures on combinatorial data. It is hard to say how big a corpus needs to be. A rather small corpus might yield good results for high-frequency words, but in general, big corpora are preferred for collocation-extraction tasks. Availability is the other key issue. Either you need already calculated co-occurrence data or full-text access to the corpus for the calculation of association measures.

Both *Korpus C4* and *DeReKo* do not meet these criteria sufficiently. Even though *Korpus C4* is is well-balanced, it is comparatively small. Besides, it does not give full-text access and offers no pre-calculated co-occurrence data. Therefore, in a lexicographical project on collocations, it can only be used for the verification of single collocations and as a source of example sentences. The problem with *DeReKo* on the other hand is not size but availability only. There is no full-text access. However, there is the possibility to calculate co-occurrences on the fly and display counts by country of origin. The disadvantage lies in the fact that this feature can only be accessed via *DeReKo*'s own graphical user interface, not via a web service or the like. This, and the on-the-fly calculation[4] make it very hard to integrate it efficiently into a lexicographical workflow where other corpora and data sources are used at the same time. For lexicographers, it is important to have quick access to all relevant data. It would be too time-consuming to manually formulate the query and wait for the calculation of the result — unless it was the only corpus used. For several reasons, it had never been an option to use *DeReKo* as the only corpus in the whole project.

Thus, we decided to build our own corpus that would provide us with data about regional variation.[5] We felt that the easiest way to get as much textual data as possible, differentiated by country at the same time, was to build a web corpus with a subcorpus for each of the three country code top-level domains *at* (Austria), *ch* (Switzerland) and *de* (Germany).

The BootCaT toolkit [5] version 0.1.5 was used to collect and post-process the web texts. For seed generation, we took around 1000 words of middle frequency from the *DeReWo* base-form list [15] and about 1400 words from a wiktionary base-vocabulary list [23]. The idea was to get texts addressed to a greater public as well as more personal texts [22, 6]. As seeds we used about 15000 randomly combined three-word sequences out of words from this list. These seeds were then used as search terms to find web pages in German — separately for each of the three country code top-level domains (*at*, *ch*, *de*). The search was conducted through the *Yahoo API*.[6] One of the major advantages of using a commercial search engine is the fact that one can profit of the search engine's mechanisms to get "good" hits, i. e. pages that are relevant for humans. For every search term, the first ten hits were stored in a URL list. That way, the three subcorpora turned out about equal in size — even if there is much more web content available under the top-level domain *de* than it is for *at* and *ch*.

All non-duplicate URLs of this list were then downloaded using the BootCaT scripts. First post-processing steps were performed with the BootCaT scripts as well. This included boilerplate stripping, filtering for html documents and language filtering with a white list of frequent function words where the texts need to contain a certain proportion of them in order to be accepted. Duplicate and near-duplicate pages were excluded using a script by Serge Sharoff[7]. All remaining web pages were transformed to a simple TEI format.

Furthermore, the corpus texts were lemmatised and part-of-speech tagged by means of the *TreeTagger* [18]. For indexing, the texts were brought into a CWB conformant one-token-per-line format and indexed with the linguistic search engine *DDC* [21]. We used a slightly modified version of the web GUI developed for the *Swiss Text Corpus* [8] to view query results and KWICs.

The final size of our web corpus amounts to 775 million tokens or about 650 million running words — distributed to the three subcorpora as follows: Austrian subcorpus (AT): 274 million tokens / 230 million running words; Swiss subcorpus (CH): 204 million tokens / 170 million running words; German subcorpus (DE): 297 million tokens / 250 million running words.

4. COLLOCATION EXTRACTION

Collocations, or more exactly, co-occurrences or collocation candidates were extracted with two different methods. On the one hand, we used traditional distance-based co-occurrences, and on the other hand, linguistically pre-processed data. Linguistic pre-processing consisted of noun chunking by *LoPar*, a parser for probabilistic context-free grammars, and the corresponding parameter file for German noun chunking [19, 20]. In this chunked data, we searched for certain patterns, such as NP + PP or attributive adjective in an NP, a procedure quite similar to that in e. g. the *SketchEngine* [16], only that the linguistic pre-processing steps are slightly different.

Distance-based co-occurrences were mainly used for noun-verb collocations. The comparably free word order of the German language makes it difficult to use pattern-based methods like the one above for noun-verb collocations. Subjects as well as objects can be found before or after the verb. To get reliable results you would need a complete syntactic analysis of the sentences. Since we felt that the performance of readily available deep parsers for German was not high enough for that purpose, we resorted to distance-based co-occurrences. To get this data, all word pairs occurring in the same sentence segment within a window of five tokens were counted.

[4] It looks efficient for an on-the-fly calculation, but a query is still much more time-consuming than with pre-calculated data.

[5] This was just one reason for the creation of this corpus, albeit an important one.

[6] Corpus creation took place in November 2010 when *Yahoo*'s search interface was still available for such purposes. Nowadays, e. g. *Bing* is a search engine that can be accessed via API.

[7] The script is called `dedupes.pl` and can be found on `http://corpus.leeds.ac.uk/tools/`. It is based on ideas from [24].

Regardless of the primary extraction method, we calculated different association measures (t-score, log-likelihood, mutual information) for the counted co-occurrences using the *UCS toolkit* [12] and ranked them according to these association measures for later lexicographical selection and processing.

5. VARIATION RECOGNITION

Comparing subcorpora of the areas in question is a straightforward method to detect regional variation in collocations. In general, if a collocation occurs in subcorpus A and not in subcorpus B, or much more frequently in subcorpus A than in B, it is assumed to be a variant of A. Not all types of variation discussed in 2.2 can be found the same way, though. Whereas collocational variants involving lexical variation get different counts in the process of collocation extraction (see section 4) syntactic and morphosyntactic variation is much more difficult to discern the way we proceed (n-grams and morphological analysis would probably be necessary for an automated recognition of this type of variation).

In this article, we will concentrate on the first type of variation, which is where the heart of collocational variation lies: otherwise Common German words that form collocations with regional variation. Inherited variation (syntactic, morphosyntactic and lexical) is secondary here.

In [14] a method is proposed to measure regional variation where relative frequencies of collocations in different subcorpora are compared to each other. This goes back to the "ratio of relative frequencies"[8] that was first used in terminological work to find specialised terms in texts of a specific field [2]. Instead of comparing a specialised corpus to a general reference corpus, two corpora from different regions are compared. A high ratio of relative frequencies of an expression shows high usage in the relevant corpus with simultaneous low usage in the second corpus.

We have adopted this method and adapted it to our needs. One question is whether to compare all of our three subcorpora to each other or to compare them each to the whole corpus as a common reference corpus. We opted for the first possibility. In our case, with three subcorpora, this leads to six values for the ratios of relative frequencies because each subcorpus is compared to the two other subcorpora. This represents a certain disadvantage to the only three values you would get from a comparison to one reference corpus. Yet, we had two cases in mind that we wanted to deal with on equal grounds: one is the "minority case" where you have a variant that is well attested in one subcorpus and less in the two others. The other case, the "majority case" is the negation of this, i. e. a variant is well attested in two subcorpora but less in the remaining one. Table 1 gives an overview of the two different modes of comparison for the model cases. With comparison of each subcorpus to each other subcorpus, minima and maxima are the same for both cases. When comparing each subcorpus to the complete corpus, minimal and maximal ratios of relative frequencies are different for the "minority" and the "majority case". Therefore, the subcorpora are all compared to each other. The maxima indicate variants (the minima non-variants respectively).

Table 1: Ratios of relative frequencies (rrf) with multiple or only one reference corpus

Corpus	rel. f	rrf$_{separate}$		rrf$_{total}$
"Minority Case"				
AT	1	1	.1	0.25
CH	1	1	.1	0.25
DE	10	10	10	2.50
AT+CH+DE	4			
"Majority Case"				
AT	10	1	10	1.43
CH	10	1	10	1.43
DE	1	.1	.1	0.14
AT+CH+DE	7			

Furthermore, the relation to the regional variation of the words forming a collocation can be taken into account using the same measures. The ratio of relative frequencies can be calculated for single words, too. Subsequently, the ratios of the collocation components can be compared to the ratio of the whole collocation, so as to see if variation is inherited from a component or rather, if it has its origin at the collocational level.

6. VERIFICATION

On a practical side, we want to look at some already known instances of regional variants in order to get a first impression on how suitable our web corpus is for questions of regional variation. We will mainly focus on collocations and other multi-word expressions for these cross-checks.

Some examples of collocations of German with heterogeneous regional distributions have been given in [14]. Actual figures based on a newspaper corpus are included there as well. An important resource for regional variants in German is the *Variantenwörterbuch des Deutschen* [4]. Its main interest lies in lexical differences between varieties of German, but it also provides a certain number of phraseologisms, and among them only a few collocations can be found. These have not been collected systematically, but the ones indicated in the dictionary should be sufficient for our purposes. As [4] is a dictionary, the frequency data the authors based their decisions on (see [9]) is not available for direct comparison. Below, we will discuss our own corpus results for some of the multi-word expressions in the two works cited [4, 14].

In [14], the collocations *tiefer Preis* (deep price, 'low price') and *einen Kredit sprechen* (a credit speak, 'to grant a credit') are cited as Helvetisms, *regierender Weltmeister* ('reigning world champion') as an Austriacism, all on a collocational level, i. e. the single component words are Common German and do not show any special regional distribution. These findings are confirmed by our web corpus where *tiefer Preis* is 55 times more frequent in the CH subcorpus than in the DE subcorpus (the factor is 175 in [14]), and 32 times more frequent than in the AT subcorpus. Compared to the DE as well as to the AT subcorpus, *einen Kredit sprechen* is about 8 times more frequent in CH. *Regierender Weltmeister* is found exclusively in the AT subcorpus (with 14 instances).[9]

[8] Cited by [14] as "weirdness ratio".

[9] For the last two examples, there are no numbers given in

Table 2: Compared ratios of relative frequencies for selected *Preis* collocations

Collocation	$\mathbf{rrf}_{ch:de}$ in [14]	$\mathbf{rrf}_{ch:de}$ Web Corpus
hoher Preis	1.2	1.2
günstiger Preis	5.2	0.8
attraktiver Preis	14.0	1.5
tiefer Preis	175.2	55.3
erschwinglicher Preis	2.0	1.2
niedriger Preis	0.3	0.7
stolzer Preis	1.7	1.7

All three examples show their regional variation on a collocational level only. None of the component words (*tief*, *Preis*, *Kredit*, *sprechen*, *regierend*, *Weltmeister*) is — according to our web corpus — itself a regional variant.

The following collocations are given in [14] as variants for Germany to the three examples above: *niedriger Preis* ('low price'), *einen Kredit bewilligen* ('to grant a credit') and *amtierender Weltmeister* ('reigning world champion'). It is not entirely clear if these are really meant as variants proper to Germany. Our data would rather suggest that they are Common German collocations or even another Helvetism in the case of *einen Kredit bewilligen*, which was found to be about 18 times more frequent in the CH subcorpus than in DE. The other two show a more or less even distribution: *niedriger Preis* and *amtierender Weltmeister* are somewhat less frequent in the CH and AT subcorpus, respectively. This lower frequency can be explained by the concurrent use of the regional variants *tiefer Preis* and *regierender Weltmeister.* But without this knowledge, and confronted to the bare frequencies, we could not assign any meaning to such relatively small differences (about 30–45% more for the DE variants).

A more systematic comparison can be found in table 2. It shows some adjective-noun collocation with *Preis*, taken from page 552 of [14]. The values in the second and third columns are the ratios of relative frequencies in the direction CH to DE in [14] (2$^{\text{nd}}$ column) and our own web corpus (3$^{\text{rd}}$ column). Very high values indicate Helvetisms, very low ones Teutonisms. In many cases the two corpora show similar results. Only for *günstiger Preis* (favorable price 'low price') and *attraktiver Preis* ('attractive price') our web corpus does not indicate a Helvetism. If we look up these two collocations in *DeReKo* we can see that they are more or less uniformly distributed over the three countries there, too — which would support our data. As to *niedriger Preis*, presented as a Teutonism, where we have just said that our corpus cannot fully support this interpretation, *DeReKo*'s distribution is similar to the one in [14] according to which *niedriger Preis* is three times more frequent in DE compared to CH.

We will now turn to examples from the *Variantenwörterbuch*[4]. Most of the marked multi-word expressions there are idiomatic phraseologisms. However, there are a few that are more compositional in meaning and consist of otherwise mostly Common German components like the examples we have already discussed. However, they do not meet these requirements as perfectly as the examples above. The collocations picked out for comparison with our web corpus are *wilder Knoblauch* AT DE ('wild garlic', Common German *Bärlauch*), *Blaulicht und Sirene* CH DE, *Blaulicht und Folgetonhorn* AT, *Blaulicht und Martinshorn* DE (variants of 'blue lights and sirens'), *in angetrunkenem Zustand* CH DE (in drunk condition 'drunk'), *Einspruch einlegen* DE ('to lodge an appeal'), *große Töne spucken* DE (spit big tones 'to talk big'). Table 3 shows the relative frequencies of these expressions in all three subcorpora, normalised to occurrences per 100 million words[10].

Table 3: Regionally marked multi-word expressions in the *Variantenwörterbuch* and their relative frequencies in the web corpus

		$\mathbf{rf}_{at}$	$\mathbf{rf}_{ch}$	$\mathbf{rf}_{de}$
wilder Knoblauch	AT DE	1.8	1.0	1.4
Blaulicht und Sirene	CH DE	2.2	5.4	3.7
Blaulicht und Folgetonhorn	AT	1.8	0.0	0.0
Blaulicht und Martinshorn	DE	1.8	1.5	5.4
in angetrunkenem Zustand	CH DE	1.1	60.3	5.0
Einspruch einlegen	DE	20.8	33.4	85.1
große Töne spucken	DE	21.2	13.2	19.2

We have all possible degrees of agreement here, ranging from support to indifference to contradiction. Whereas *große Töne spucken* and *wilder Knoblauch* rather look like Common German in our data, and not DE or non-CH as suggested by [4], the areal markings in the dictionary for *Blaulicht und Folgetonhorn*, *Blaulicht und Martinshorn* and *in angetrunkenem Zustand* are clearly supported by the webcorpus data. To a lesser degree, this is also valid for *Einspruch einlegen* — although it is attested in AT and CH, it is much more frequent in DE. For *Blaulicht und Sirene* our data is not very clear. The trend suggests the markings CH and DE, but AT is well attested, too.

If we compare these findings to *DeReKo* we get a picture similar to our web corpus. One difference is that distinctions have a slight tendency to be more clear-cut in *DeReKo.* Thus, according to *DeReKo* the collocation *Blaulicht und Sirene* is an obvious non-Austriacism. Besides, the ratios of relative frequencies often map the ones in table 3 rather well. E. g. *in angetrunkenem Zustand* with a distribution of 1:60:5 (AT:CH:DE) in the web corpus shows a distribution of approximately 1:40:15 in *DeReKo.* The whole picture is not completely consistent. If a trend can be observed, then it is the one that in the web corpus DE variants are sometimes more frequent in the AT and CH subcorpora than they are in the *DeReKo* AT and CH parts. Several reasons could have led to this. It is much more probable that a text from Germany might find its way to a website in the *ch* top-level domain than it is for the same text to get into a Swiss newspaper. Moreover, people might tend to write for a greater public and avoid regional variants when they write for the web (and the German from Germany might still be for many people the "non-local" variety). Nonetheless, the web corpus still shows clear distinctions and approximately the same distinctions between the three countries as *DeReKo.*

[14], hence the missing comparison.

[10] Space-separated tokens, to be exact.

7. WEB CORPUS RESULTS

After verification and cross-comparison with findings originally based on other corpora and methods, this section describes the usage of the web corpus in lexicographical work where it serves to find new regional variants of collocations. Moreover, the collected corpus data can provide a first idea of the extent and the kind of regional variation that collocations exhibit.

7.1 Lexicographical Work

First, the question arises whether our web corpus alone is sufficient to decide about regional variants on a collocational level. When variation is inherited from the lexical level, we have dictionaries to our aid (e. g. [4]), and we have much easier access to corpus data for single words than for collocations. Furthermore, purely collocational variants are often not salient. Contrary to many single words, speakers (and lexicographers) tend not to have conscious knowledge of the regional distribution of collocations. This makes the task more difficult. From our experience, a quick or more extended look at the actual text samples is nearly always needed in order to confirm or reject the mere figures.

There are a number of problems we observe when working with our web corpus on regional variation. One major problem is duplicated text. The measures that have been taken against it (boilerplate stripping, duplicate and near-duplicate detection, see above) could still be reinforced and enhanced. However, when looking at the KWICs, these instances can be identified at a glance because of the regular patterns they form. Increased efforts on duplicate elimination would certainly prove beneficial to the accuracy of the variation figures. However, as long as the KWICs have to be consulted for other reasons, too, the remaining text duplicates are not very time-consuming, though annoying.

Another source of errors is differing corpus composition in general. This is a problem that one will probably always encounter when comparing corpora. Even though we have tried to minimise differences between subcorpora by using the same methods, the same search terms and making the subcorpora all approximately the same size, there are still differences that are difficult to explain. It is often not obvious if a difference mirrors a social or cultural fact (including main news topics that differ from country to country) or if it is just due to random text choice. One example in our data is the collocation *Milch aufkochen* 'to boil up milk' that is attested with 5 to 7 times higher frequency in AT and CH than in DE. In *DeReKo*, the same collocation is more frequent in AT and less in CH and DE. Finally, if we compare to *DWDS* (core corpus, a reference corpus for Germany), we get a much higher relative frequency than for any of the subcorpora of our web corpus. The collocation *Milch aufkochen* is frequently used in recipes. The default suspicion would then be that the frequency of this collocation depends on the frequency of recipes in the corpus, hence no regional variation. Yet, based on this data, it is not possible to exclude that it might be used much more often in e. g. Austrian recipes.

Another difficulty is data sparseness: for the many collocations with low overall frequencies — an exact threshold is hard to establish — it is often impossible to state anything about their regional distribution. A collocation may display a frequency of e. g. five occurrences in the whole corpus, distributed 4:1:0 onto the three subcorpora. Without any additional information, it is impossible to decide whether the four occurrences in one subcorpus indicate a regional variant or whether it is rather coincidence.

Furthermore, the status of the ratio of relative frequencies is not always very clear or cannot serve as the only value to measure regional variation. In a situation of variation when variants are compared to "non-variants", typically, high numbers are divided by low numbers. Especially for the low numbers, noise and errors such as the ones mentioned above cause dramatical changes in the ratio of relative frequencies. On the other hand, when the lower frequency itself is already quite high in absolute terms[11], it is often difficult to speak of a "non-variant" because the collocation is undeniably well attested. In these cases, the ratio does not exactly indicate where to draw a border between "absolute variants" and "variants of frequency"[12].

Despite all these shortcomings, our web corpus is still a powerful means to detect regional variants of collocations. Most of the time, it is in line with other corpora such as *DeReKo* and when it is not, this can often be easily noticed (e. g. duplicates). The major advantage for us is that thanks to our web corpus, we get systematic data about the regional distribution of collocations and can integrate it fully into our dictionary management system. Without it, we might have overlooked many regional variants that are not so obvious otherwise, like e. g. *Auskunft über die Durchführung* CH 'information about the carrying out' – *Kost und Quartier* AT 'board and lodging' – *ungenügende Leistung* CH 'insufficient performance' – *Gemüse putzen* AT, DE 'to prepare vegetables for cooking' – *den Fernseher anschalten* DE 'to switch on the television' – *verlängertes Wohnzimmer* AT 'extended living room' – *Höhe über Meer* CH 'height above sea level'.

7.2 Variation Distribution

Another important point is to know how much regional variation in German collocations there actually is. Where variation is inherited from other levels such as lexicon and syntax, the situation is better inasmuch as there has already been considerable research in these fields (for the lexicon: [4]; for syntax: [11]). Concerning collocations proper, it is very unclear how much regional variation there is on a purely collocational level, i. e. without the effects of lexical or syntactic variation. Guesses differ widely. However, our data allows for certain quantitative statements here.

Figure 1 shows maximal ratios of relative frequencies for a random sample of collocations in our dictionary. Each point represents one collocation with its maximal ratio of relative frequencies vertically (out of the six possible values) and the maximal ratio of relative frequencies of its component words horizontally.[13] Or, in other words, you have the overall vari-

[11] Or in relation to one of the components.

[12] Such variants of frequency would be marked with *especially XY* — a marking we had excluded in our dictionary project.

[13] Only collocations with a minimal absolute frequency of 50 have been considered here. For the chart, 0 frequencies have been smoothed using an "add-1" method in order to prevent division by 0 (this equally applies to the data in figure 2).

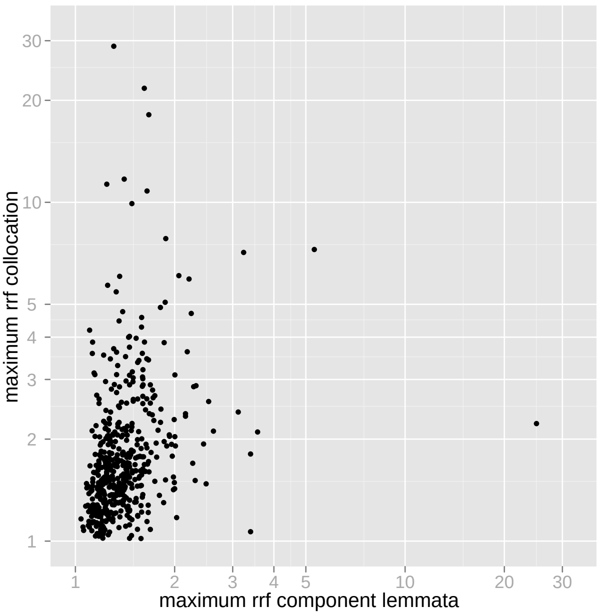

Figure 1: Maximal ratios of relative frequencies of collocations and their components

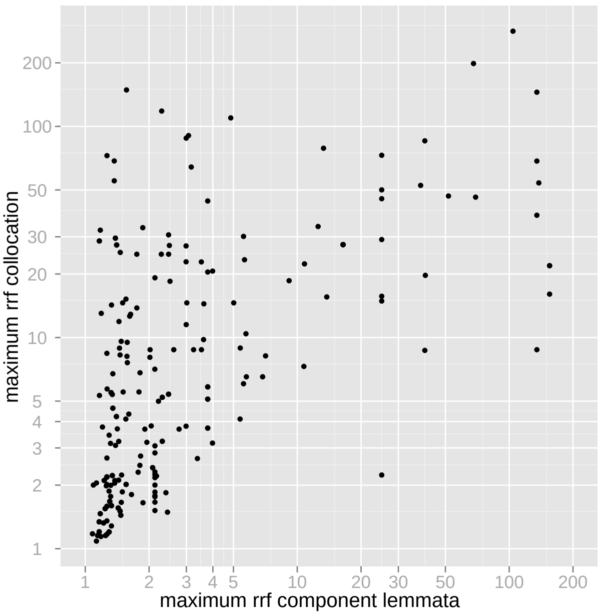

Figure 2: Maximal ratios of relative frequencies of collocations with regional markings and their components

ation score of the collocation itself (vertically) against the variation score inherited from its component words (horizontally). Collocations with regional variation tend to be towards the top, the ones with the variation inherited from their components tend to be towards the right side, others with inconspicuous component words are found at the left side.

What can be derived from the data displayed in figure 1 is that regional variation in German collocations is by no means just inherited from lexical variation of the component words. Variation on a purely collocational level is at least as frequent or even more frequent if we look at the data points just above the main "point cloud" on the left. However, its true extent is difficult to estimate. Cases in which one component word is Common German but a regional variant in one of its meanings cannot be distinguished with our data. Moreover, the data does not suggest a natural limit above which something has to be a regional variant, but rather shows the gradient nature of the phenomenon.

In order to see what kind of relative-frequency-ratio patterns regional variants may form, we can have a look at collocations that have already been marked as regional variants by lexicographers. Figure 2 shows the same values as figure 1, but for collocations marked as regional variants only. The data points are more or less uniformly distributed over the area above the diagonal from the lower left to the upper right corner of the chart, as might have been expected.

This supports the ratio of relative frequencies as a sensible measure of regional variation. Moreover, it also supports the assumption discussed above that regional variation in German collocations is not only inherited from its lexical components but occurs to about the same extent on a more or purely collocational level.

Further research is needed if one wants to distinguish between "purely collocational" and "lexical meaning-related" variation, and in order to measure their respective quotas. Another open question demanding more research is if a limit can be named for a ratio of relative frequencies above which the chances of facing a regional variant are very high.

8. CONCLUSION

We have shown how a web corpus can be used to recognise regional variation in German collocations for the three main national varieties (Austria, Germany, Switzerland). It is a useful tool, on a level with other corpora suitable for the same purpose. Advantages such as easy text collection and corpus creation, full-text access and the possibility of full integration into the dictionary management system outperform disadvantages like duplication problems and partly uneven text distribution. If web corpora can also be used for the distinction of more fine-grained regional variation than by countries is subject to further research.

The web-corpus data suggests that regional variation in German collocations is as frequent on a purely collocational level (including lexical meaning-related variation) as it is as a result of lexical variation and, as it were, inherited. Further research is needed to establish a more detailed view of these relations.

9. REFERENCES

[1] A. Abel, S. Anstein, and S. Petrakis. Die Initiative Korpus Südtirol. *Linguistik online*, 38(2):5–12, 2009.

[2] K. Ahmad, A. Davies, H. Fulford, and M. Rogers. What is a term? The semi-automatic extraction of

terms from text. In M. Snell-Hornby, F. Pöchhacker, and K. Kaindl, editors, *Translation Studies: An Interdiscipline*, Benjamins translation library, pages 267–278. John Benjamins, 1994.

[3] U. Ammon. *Die deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten.* Walter de Gruyter, Berlin, New York, 1995.

[4] U. Ammon et al. *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol.* Walter de Gruyter, Berlin, 2004.

[5] M. Baroni and S. Bernardini. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*, 2004.

[6] M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation*, 43(3):209–226, 2009. http://clic.cimec.unitn.it/marco/research.html.

[7] H. Biber, E. Breiteneder, and K. Moerth. The Austrian academy corpus – digital resources and textual studies. In *Proceedings of the 14th Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2002), Tübingen, Germany*, pages 16–17, 2002.

[8] H. Bickel, M. Gasser, L. Hofer, and C. Schön. Das Schweizer Textkorpus. *Linguistik online*, 39(3):5–31, 2009.

[9] H. Bickel and R. Schmidlin. Ein Wörterbuch der nationalen und regionalen Varianten der deutschen Standardsprache. *Bulletin VALS-ASLA (Vereinigung für angewandte Linguistik in der Schweiz)*, 79:99–122, 2004.

[10] H. Dittmann, M. Ďurčo, A. Geyken, T. Roth, and K. Zimmer. Korpus C4 – a distributed corpus of German varieties. In T. Schmidt and K. Wörner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*, Hamburg Studies in Multilingualism (HSM). Benjamins, Amsterdam, 2012. To appear.

[11] C. Dürscheid, S. Elspaß, and A. Ziegler. Grammatische Variabilität im Gebrauchsstandard – das Projekt "Variantengrammatik des Standarddeutschen". In M. Konopka, J. Kubczak, C. Mair, F. Štícha, and U. H. Waßner, editors, *Grammatik und Korpora 2009 / Grammar & Corpora 2009*, pages 123–140, Tübingen, 2011. Narr.

[12] S. Evert. *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* PhD thesis, Universität Stuttgart, Stuttgart, 2005.

[13] A. Geyken. The DWDS corpus: a reference corpus for the German language of the twentieth century. In C. Fellbaum, editor, *Idioms and Collocations*, pages 23–40. Continuum, London, 2007.

[14] U. Heid. Korpusbasierte Beschreibung der Variation bei Kollokationen: Deutschland – Österreich – Schweiz – Südtirol. In S. Engelberg, A. Holler, and K. Proost, editors, *Sprachliches Wissen zwischen Lexikon und Grammatik. Jahrbuch des Instituts für Deutsche Sprache 2010*, pages 533–558, Berlin, 2011. De Gruyter.

[15] Institut für deutsche Sprache. Korpusbasierte Wortgrundformenliste DeReWo, v-30000g-2007-12-31-0.1, mit Benutzerdokumentation, 2007. http://www.ids-mannheim.de/kl/derewo/.

[16] A. Kilgarriff, P. Rychlý, P. Smrž, and D. Tugwell. The sketch engine. In *Proceedings Euralex 2004, Lorient, France*, pages 105–116, 2004.

[17] M. Kupietz, C. Belica, H. Keibel, and A. Witt. The German reference corpus DeReKo: A primordial sample for linguistic research. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *LREC*. European Language Resources Association, 2010.

[18] H. Schmid. Improvements in part-of-speech tagging with an application to German. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D. Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, pages 13–26, Dordrecht, 1999. Kluwer Academic Publishers.

[19] H. Schmid. LoPar. Design and implementation. Technical report, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 2000.

[20] H. Schmid and S. Schulte im Walde. Robust German noun chunking with a probabilistic context-free grammar. In *Proceedings of the 18th conference on Computational linguistics, Saarbrücken, Germany*, pages 726–732, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[21] A. Sokirko. *User Manual for DWDS/Dialing Concordance*, 2005.

[22] M. Ueyama. Evaluation of japanese web-based reference corpora: Effects of seed selection and time interval. In M. Baroni and S. Bernardini, editors, *Wacky! Working Papers on the Web as Corpus*, pages 99–126, Bologna, 2006. Gedit.

[23] Wiktionary. Grundwortschatz plus Wiktionary minimum, 2008. http://de.wiktionary.org/wiki/Wiktionary:Projekt:Grundwortschatz_plus_Wiktionary_minimum.

[24] R. Ziai and N. Ott. *Web as Corpus Toolkit: User's and Hacker's Manual.* Lexical Computing Ltd., Brighton, UK, 2005. Manual for version pre3.

Efficient Web Crawling for Large Text Corpora

Vít Suchomel
Natural Language Processing Centre[*]
Masaryk University, Brno, Czech Republic
xsuchom2@fi.muni.cz

Jan Pomikálek
Lexical Computing Ltd.
xpomikal@fi.muni.cz

ABSTRACT

Many researchers use texts from the web, an easy source of linguistic data in a great variety of languages. Building both large and good quality text corpora is the challenge we face nowadays. In this paper we describe how to deal with inefficient data downloading and how to focus crawling on text rich web domains. The idea has been successfully implemented in SpiderLing. We present efficiency figures from crawling texts in American Spanish, Czech, Japanese, Russian, Tajik Persian, Turkish and the sizes of the resulting corpora.

Categories and Subject Descriptors

I.2.7 [**Computing Methodologies**]: Artificial Intelligence—*Natural Language Processing*

Keywords

Crawler, web crawling, corpus, web corpus, text corpus

1. INTRODUCTION

Most documents on internet contain data not useful for text corpora, such as lists of links, forms, advertisement, isolated words in tables, and other kind of text not comprised of grammatical sentences. Therefore, by doing general web crawls, we typically download a lot of data which gets filtered out during post-processing. This makes the process of web corpus collection inefficient.

To be able to download large collections of web texts in a good quality and at a low cost for corpora collection managed by SketchEngine[1], we developed SpiderLing—a web spider for linguistics. Unlike traditional crawlers or web indexers, we do not aim to collect all data (e.g. whole web domains). Rather than that we want to retrieve many documents containing full sentences in as little time as possible.

We have experimented with using third party software for obtaining text documents from the web. Following the example of other researchers [2, 3, 1], we have used Heritrix crawler[2] and downloaded documents for the language in interest by restricting the crawl to national web domains of the countries where the language is widely used (e.g. `.cz` for Czech). Though we managed to compile corpora of up to 5.5 billion words in this way [6], we were not satisfied with the fact that we need to keep the crawler running for several weeks and download terabytes of data in order to retrieve a reasonable amount of text. It turned out that most downloaded documents are discarded during post-processing since they contain only material with little or no good quality text.

2. ANALYSIS OF PREVIOUS WORK

We were interested to know how much data we download in vain when using Heritrix and if the sources which should be avoided can be easily identified. In order to get that information we analyzed the data of a billion word corpus of European Portuguese downloaded from the `.pt` domain with Heritrix. For each downloaded web page we computed its yield rate as

$$yield\ rate = \frac{final\ data}{downloaded\ data}$$

where *final data* is the number of bytes in the text which the page contributed to the final corpus and *downloaded data* is simply the size of the page in bytes (i.e. the number of bytes which had to be downloaded). Many web pages have a zero yield rate, mostly because they get rejected by a language classifier or they only contain junk or text duplicate to previously retrieved text.

We grouped the data by web domains and computed a yield rate for each domain as the average yield rate of the contained web pages. We visualized this on a scatterplot which is displayed in Fig. 1. Each domain is represented by a single point in the graph.

It can be seen that the differences among domains are enormous. For example, each of the points in the lower right corner represents a domain from which we downloaded more than 1 GB of data, but it only yielded around 1 kB of text. At the same time, there are domains which yielded more than 100 MB of text (an amount higher by five orders of magnitude) from a similar amount of downloaded data. These

[*] `http://nlp.fi.muni.cz/`

[1] `http://sketchengine.co.uk/`

[2] `http://crawler.archive.org/`

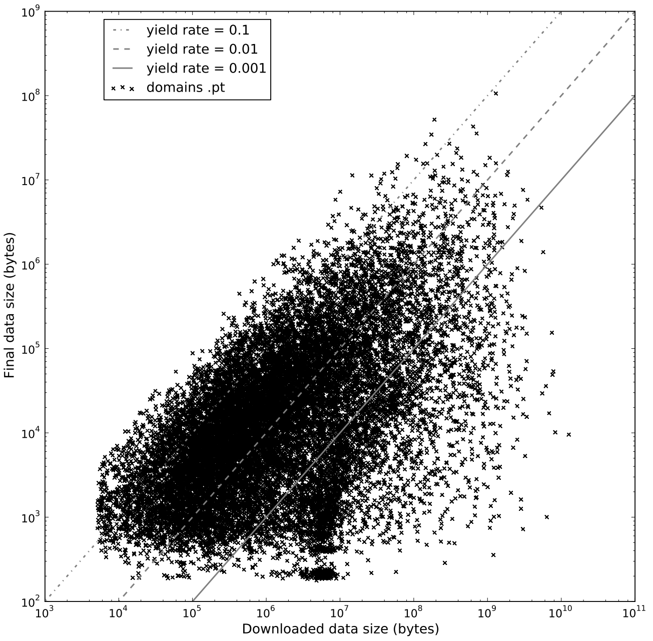

Figure 1: Web domains yield rate for a Heritrix crawl on .pt

Table 1: Sums of downloaded and final data size for all domains above the given yield rate threshold

yield rate threshold	domains above the threshold	crawler output size [GB]	final data size [GB]	final yield rate
none	51645	1288.87	4.91	0.0038
0	31024	1181.56	4.91	0.0042
0.0001	29580	705.07	4.90	0.0069
0.0002	28710	619.44	4.89	0.0079
0.0004	27460	513.86	4.86	0.0095
0.0008	25956	407.30	4.80	0.0118
0.0016	24380	307.27	4.68	0.0152
0.0032	22325	214.18	4.47	0.0209
0.0064	19463	142.38	4.13	0.0290
0.0128	15624	85.69	3.62	0.0422
0.0256	11277	45.05	2.91	0.0646
0.0512	7003	18.61	1.98	0.1064
0.1024	3577	5.45	1.06	0.1945
0.2048	1346	1.76	0.54	0.3068
0.4096	313	0.21	0.10	0.4762

domains are positioned in the upper right corner of the graph.

Next, we selected a set of yield rate thresholds and computed for each threshold the number of domains with a higher yield rate and the sum of downloaded and final data in these domains. The results can be found in Table 1.

It is easy to see that as the yield rate threshold increases the size of the downloaded data drops quickly whereas there is only a fairly small loss in the final data. This suggests that by avoiding the domains with low yield rate a web crawler could save a lot of bandwidth (and time) without making the final corpus significantly smaller.

For instance if only domains with a yield rate above 0.0128 were crawled, the amount of downloaded data would be reduced from 1289 GB to 86 GB (to less than 7 %) while the size of the final data would only drop from 4.81 GB to 3.62 GB (73.7 %). This is of course only a hypothetical situation, since in practice one would need to download at least several pages from each domain in order to estimate its yield rate. Nevertheless, it is clear that there is a lot of room for making the crawling for web corpora much more efficient.

We observe many web domains offer documents of a similar type. For example, a news site contains short articles, a blog site contains blog entries, a company presentation site contains descriptions of the goods sold or products manufactured. We believe the quality of several documents (with regard to building text corpora) on such sites could represent the quality of all documents within the given domain.

One could argue that a segmentation by domains is too coarse-grained since a domain may contain multiple websites with both high and low yield rates. Though we agree, we believe that identifying more fine-grained sets of web pages (like a text rich discussion forum on a text poor goods presentation site) introduces further complications and we leave that for future work.

3. SPIDERLING

Simple web crawlers are not robust enough to suit our needs (e.g. not supporting heavily concurrent communication, lacking load balancing by domain or IP address, not able to restart the crawling after a system crash). On the other hand, the source code of sophisticated crawlers is too complex to alter, making implementation of our way of efficient web traversing difficult.

We came to the conclusion that the easiest way of implementing our very specific requirements on web crawling is to create a custom crawler from scratch. In order to reduce the amount of unwanted downloaded content, the crawler actively looks for text rich resources and avoids websites containing material mostly not suitable for text corpora. Our hope was that by avoiding the unwanted content we can not only save bandwidth but also shorten the time required for data postprocessing and building a web corpus of given size.

3.1 Improving the yield rate

Our primary aim is to identify high-yielding domains and to avoid low-yielding ones. At the same time we want to make sure that we do not download all data only from a few top-yielding domains so that we achieve a reasonable diversity of the obtained texts.

We collect information about the current yield rate of each domain during crawling the web. If the yield rate drops below a certain threshold we blacklist the domain and do not download any further data from it. We define a minimum

Table 2: The yield rate threshold as a function of the number of downloaded documents

documents count	yield rate threshold
10	0.00
100	0.01
1000	0.02
10000	0.03

amount of data which must be retrieved from each domain before it can be blacklisted. Current limit is 8 web pages or 512 kB of data, whichever is higher. The yield rate threshold is dynamic and increases as more pages are downloaded from the domain. This ensures that sooner or later all domains get blacklisted, which prevents overrepresentation of data from a single domain. Nevertheless, low-yielding domains are blacklisted much sooner and thus the average yield rate should increase.

The yield rate threshold for a domain is computed using the following function:

$$t(n) = 0.01 \cdot (\log_{10}(n) - 1)$$

where n is the number of documents downloaded from the domain. The function is based partly on the authors' intuition and partly on the results of initial experiments. Table 2 contains a list of thresholds for various numbers of downloaded documents.

We experimented with various parameters of the yield rate threshold function. Fig. 2 shows how the average yield rate changes in time with different yield rate threshold functions. These experiments have been performed with Czech as the target language. It can be seen that stricter threshold functions result in higher average yield rate. However, too high thresholds have a negative impact on the crawling speed (some domains are blacklisted too early). It is therefore necessary to make a reasonable compromise.

Note: We used the threshold functions from Fig. 2 in our initial experiments. We selected an even less strict one (defined in this section) later on during crawling various data sources. It was a matter of balancing high yield rate versus total amount of obtained data. Too much data was thrown away due to a strict threshold. That is why the currently used threshold function is not present in the figure. The main point is that yield rate is strongly affected by the selected threshold function.

3.2 Removing junk and duplicates

We use jusText[3] [5]—a heuristic based boilerplate removal tool—embedded in SpiderLing to remove content such as navigation links, advertisements, headers and footers from downloaded web pages. Only paragraphs containing full sentences are preserved.

Duplicate documents are removed at two levels: (i) original form (text + HTML), and (ii) clean text as produced by jusText. Two correspondent checksums are computed for each web page and stored in memory. Documents with previously seen checksums are discarded. Both kinds of removal are done on-the-fly during the crawling to immediately propagate the currently crawled documents' yield rate into the corresponding domain yield rate. This enables SpiderLing to dynamically react to obtained data.

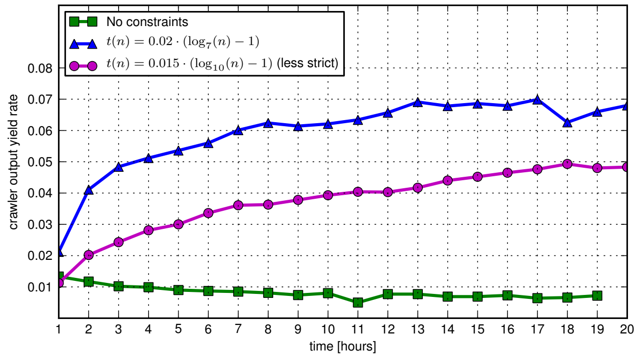

Figure 2: Average yield rate in time for various yield rate threshold functions (crawling the Czech web)

As a post-processing step, we also remove near-duplicates using onion[4]. The deduplication is performed on paragraph level. Paragpraphs consisting of more than 50 % word 7-tuples encountered in previously processed data are removed. Since such deduplication is a higly demanding task in terms of both processor cycles and memory consumption, we did not embed it into SpiderLing. Nonetheless, we are still considering such integration, since it would enable a more accurate estimate of yield rates and thus improve the crawler's traversing algorithm.

We currently do not filter unwanted web content such as link farms and machine generated texts. This may be a subject to further research. Note though that some of such content (e.g. excerpts of Wikipedia articles on link farms) is already reduced in our current processing pipeline as a positive side effect of deduplication.

4. RESULTS

4.1 Created corpora

During the development of the crawler we downloaded a total of ca. 4 TB Czech web pages in several web crawler runs. This amounts to ca. 5 billion tokens after all post-processing steps, including deduplication with onion. We merged the corpus with a ca. 2 billion word Czech web corpus we have collected previously by using Heritrix. Since the two corpora overlapped to a high extent, the size of the final Czech web corpus after deduplication is 5.8 billion tokens.

As a next exercise we ran SpiderLing on Tajik Persian to find out how the crawler deals with scarce online resources. We started the crawl from 2570 seed URLs (from 475 distinct domains) collected with Corpus Factory [4]. The crawling finished in two days having no more URLs to download from.

[3] http://code.google.com/p/justext/

[4] http://code.google.com/p/onion/

Table 3: Final corpora sizes obtained using SpiderLing

target language	corpus size [10^9 tokens]	crawling duration [days]
American Spanish	8.7	14.0
Arabic	6.6	28.0
Czech	5.0	24.4
Japanese	11.1	28.0
Russian	20.2	13.4
Tajik Persian	0.034	1.7
Turkish	3.1	7.0

Since then we focused on crawling widely spread languages such as American Spanish, Japanese and Russian. There are many resources in those languages available on the web, which contributed to quick and effortless crawling.
Since the crawler supports constraining by internet top level domain, the American Spanish corpus was downloaded from national domains of 18 Hispanic American countries. Documents from the same country form a subcorpus of the resulting corpus. Such data may be useful for a research studying varieties in the Spanish language spoken in America. It is worth noting that 75 % of documents in the corpus come from three countries with the highest web presence: Argentina, Mexico and Chile.

There is an overview of all corpora recently built with SpiderLing in Table 3.

4.2 Yield rate

By applying yield rate thresholds on domains we managed to reduce downloading data which is of no use for text corpora and increased the overall average yield rate. Fig. 3 contains the same kind of scatterplot as displayed in Fig. 1, this time on the data downloaded by SpiderLing with Czech as a target language. This is a significant improvement over the previous graph. For low-yielding domains only up to 1 MB of data is downloaded and high amounts of data are only retrieved from high-yielding sources. Many points (i.e. domains) are aligned along the line representing a yield rate of 10 %. Furthermore, the crawling was stopped already at the 512 kB threshold in case of many bad domains.

Note that the graph in Fig. 3 does not take deduplication by onion into account. It displays the size of the data as output by the crawler (i.e. boilerplate removed by jusText, no exactly same documents), not the final deduplicated texts size. Even though the achieved improvement over the previous is indisputable.

We were also interested in the development of the crawling efficiency during crawling. We expected the yield rate to slightly increase over time (the more data downloaded the higher yielding domains selected). The results are pictured by Fig. 4.

Contrary to our expectations, the measured efficiency grew only slightly or stagnated in most cases. We still consider this a good result because even the stagnating yield rates

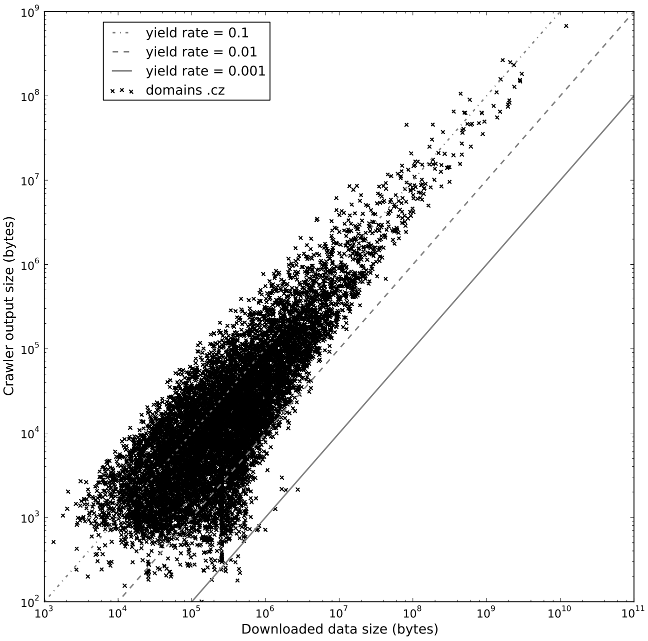

Figure 3: Web domains yield rate for a SpiderLing crawl on the Czech web

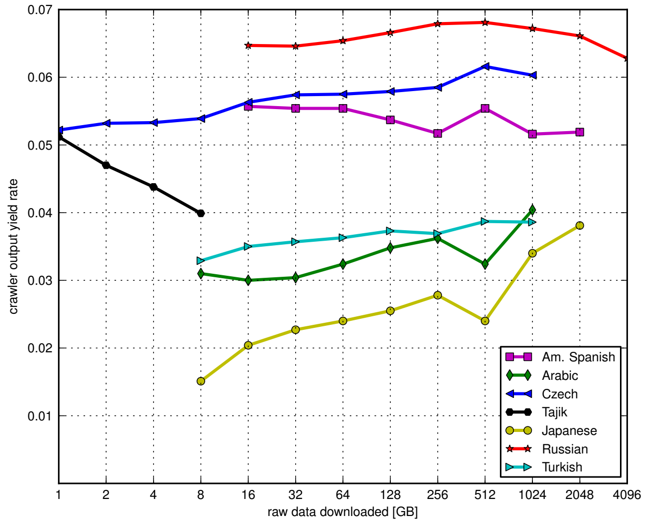

Figure 4: Web domains yield rate for SpiderLing crawls on six target languages

Table 4: Results of crawling the web for large text corpora using SpiderLing

target language	raw data size [GB]	crawler output size [GB]	crawler output yield rate	final corpus size [GB]	final yield rate
Am. Spanish	1874	97.3	0.0519	44.14	0.0236
Arabic	2015	84.4	0.0419	58.04	0.0288
Japanese	2806	110.1	0.0392	61.36	0.0219
Russian	4142	260.1	0.0628	197.5	0.0477
Turkish	1298	51.4	0.0396	19.52	0.0150

were good (with regard to Table 1).
Crawling Japanese was an exception, since the rate kept increasing almost all the time there. The reason may be the starting rate was low. The inbuilt language dependent models (character trigrams, wordlist) may not be adapted well for Japanese and throw away good documents as well. The less web resources in the target language, the sooner the yield rate drops down. It can be demostrated by the example of Tajik.

The initial yield rate obviously depends on the quality of the seed (initial) URLs. (For example many URLs of electronic newspaper articles in the target language give good initial yield rate.) Irrespective of the seed URLs, the measurements show that sooner or later, the program discovers enough URLs to be able to select good quality domains.

Unlike other languages, crawling Arabic, Japanese and Turkish was not restricted to the respective national domains. That inevitably led to downloading more data in other languages thus throwing away more documents. Considering crawling efficiency in these cases on Fig. 4, the yield rate also depends on constraining crawling to national top level domains.

The yield rate may decrease after downloading a lot of data (the amount depends on the web presence of the target language). In the case of rare languages, the best (text rich) domains get exhausted and the crawler has to select less yielding domains. Concerning decreasing rate while crawling widely spread languages (like Russian), the reason may lie in the design of the crawler. It may be obtaining data from many domains concurrently (ca. 1000–2000 in case of Russian), leaving potentially rich domains waiting and not discovering their true potential.

The final data sizes and average yield rates obtained by crawling five large languages are summarized in Table 4. The final yield rate varies between 1.50 % (Turkish) and 4.77 % (Russian) which is a great improvement over a yield rate of 0.38 % achieved by Heritrix (crawling Portuguese) and a good result compared to a hypothetical yield rate of 1.28 % discussed in section 2.

5. FUTURE WORK

We plan building more huge corpora covering all major languages (French and English being next on the list). Since there are many online resources in these languages, we expect to gather two at least 20 billion tokens corpora in less than a month.
We also need to invest more effort into optimizing the program design to gain more data from scarce resources.
We would like to interconnect the crawler with other linguistic and data processing tools to form a single system offering instant web corpora on demand.

Other plans for the future include analyzing the topics and genres of the downloaded texts and eventually balancing the downloaded content in this respect.

6. CONCLUSION

We presented a way of selective crawling to make obtaining internet content for text corpora more efficient. We have implemented the idea in a new web crawler, which can effectively avoid data not suitable for text corpora thus significantly improving the yield rate of the downloaded documents.

The crawler has already been successfully applied for building billions of words scale corpora in six languages. Texts in the Russian corpus, consisting of 20.2 billions tokens, were downloaded in just 13 days.

7. REFERENCES

[1] M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009.

[2] M. Baroni and A. Kilgarriff. Large linguistically-processed web corpora for multiple languages. *Proceedings of European ACL*, 2006.

[3] A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (LREC 2008)*, 2008.

[4] A. Kilgarriff, S. Reddy, J. Pomikálek, and A. PVS. A corpus factory for many languages. *Proc. LREC, Malta*, 2010.

[5] J. Pomikálek. *Removing Boilerplate and Duplicate Content from Web Corpora.* PhD thesis, Masaryk University, Brno, 2011.

[6] J. Pomikálek, P. Rychlý, and A. Kilgarriff. Scaling to billion-plus word corpora. *Advances in Computational Linguistics*, 41:3–13, 2009.

Not Just Bigger: Towards Better-Quality Web Corpora

Yannick Versley
SFB 833
Universität Tübingen
versley@sfs.uni-tuebingen.de

Yana Panchenko
Seminar für Sprachwissenschaft
Universität Tübingen
yana.panchenko@uni-tuebingen.de

ABSTRACT

For the acquisition of common-sense knowledge as well as as a way to answer linguistic questions regarding actual language usage, the breadth and depth of the World Wide Web has been welcomed to supplement large text corpora (usually from newspapers) as a useful resource.

While purists' criticism on unbalanced composition or text quality is easily shrugged off as unconstructive, empirical results in some real-world tasks have found Web corpora to be less useful than (smaller) newspaper corpora. More than the early criticism, evidence that Web corpora are doing poorly at their original purpose should raise concerns about the quality of Web corpora. Especially for non-English Web corpora, principled quality assessment and targeted improvements are instrumental in ensuring their relevance.

In this paper, we present our own pipeline for Web corpora, which includes improvements regarding content-sensitive boilerplate detection as well as language filtering for mixed-language documents. We also provide a principled evaluation of the combination of corpora and (non-linguistic and linguistic) preprocessing between more standard types of large corpora (newspaper and Wikipedia) and different Web corpora.

While our current results are focused on German-language Web corpora, both the content-sensitive boilerplate detection and our method of evaluation by constructing an artificial thesaurus from a wordnet are applicable to many other languages.

1. INTRODUCTION

Unsupervised and semi-supervised learning methods use vast quantities of text to improve the coverage and accuracy of language processing over that possible with only hand-annotated data. Large collections of suitable text are important for such learning methods: They extract features from the context of target items, and a large number of contexts (i.e., a large number of texts) is instrumental in avoiding sparse data problems in the learning task(s).

Many kinds of text are used for distributional semantics – large fixed text collection such as newspaper text, Wikipedia or the Gutenberg project, n-gram frequency databases extracted from larger text collections, or online queries to search engines. Among these types, however, Web corpora are the only type that combines replicability of results (online queries may yield different results), scalability in size (unlike single-source text collections) as well as the possibility of using rich linguistic annotations (unlike n-gram databases).

In some cases, however, Web corpora seem to be less useful despite their larger size: [BL08] found that the much smaller British National Corpus (BNC; 100 million tokens) was better-suited for the creation of a window-based distributional similarity resource than the much larger ukWaC Web corpus ([BBFZ09]; 2.25 billion tokens). For their more targeted approach using patterns, the larger size of the Web corpus together with the filtering ability of their learning algorithm were able to make better use of the greater size of ukWaC.

[FP10] used large corpora in the context of building a Named Entity Recognizer (NER) system for German; in their case, they induce a word clustering on the unannotated corpus – either 175 million tokens of German newspaper text (HGC), or the same amount of text from the deWaC Web corpus. They found that the HGC-derived word clusters were substantially more useful in the case of in-domain testing, and somewhat more useful in the case of out-of-domain testing data containing parliamentary debates.

It would be hardly surprising in general that newspaper text is a better source of in-domain unannotated textual data than general-domain text from the World Wide Web. However, the advantage of the (cleaner) newspaper text persists to a certain extent when testing on out-of-domain texts from European Parliament debates. Similarly, the balanced composition of the British National Corpus does not seem any less heterogeneous or a more straightforward match to Baroni and Lenci's psycholinguistic data than the ukWaC corpus counterpart.

These results should be seen in contrast to the evaluation by [LC06], who also compare newspaper text with a Web corpus. On a distributional semantics task, they show a relatively small difference in scores for size-matched two billion word samples of the English Gigaword Corpus[1] and their own Web corpus.

In sum, these results make it seem worthwile to investigate the factors behind the sensitivity to differences between Web and newspaper corpora. Several attributes – text selection and domain distribution, text quality, the preprocessing used – may contribute to this sensitivity.

In terms of **domain distribution**, the World Wide Web contains texts from a large range of domains, with a few frequent domains that dominate the statistics extracted from such text collections: [Ver08] finds that the largest singular vectors in an SVD model induced from bigrams in Google's English ngram dataset (containing a verb as the first word) reflect predominant domains, including product sales, legal texts, pornography, Unix manuals, and news stories.

The distribution of **text genres** of Web text – how a given topic is talked about, as opposed to the topic itself – is also markedly different from the genres found in newspaper or newswire text. In principle, this could be seen as beneficial - after all, the use of edited newspaper text for knowledge acquisition was motivated by availability rather than other reasons, and informal writing on the Web may be closer in genre to the language use of ordinary people. More importantly, however, informal or unedited text is more difficult to process due to variation in spelling and punctuation, or the presence of more grammar deviations.

The greater variability of Web content also leads to other quality issues resulting from greater variability in how the text is presented: While removing formatting information from single-source text is usually feasible to do in near-perfect quality, Web corpora have to rely on generic techniques for **boilerplate removal** that use effective heuristics (e.g., detecting navigation elements based on the length of the text between two HTML tags, cf. subsection 2.3) but are not perfect.

Last but not least, **preprocessing tools** such as the tokenizers or part-of-speech taggers used on Web corpora may yield lower quality - either because the best-quality processing tools cannot be used for speed or licensing reasons, or because the domain and genre distribution of Web corpora makes the processing of such text more difficult.

In the remainder of this paper, we adopt a similar evaluation framework to that of [LC06] in comparing several large German corpora from a distributional semantics perspective that show different domain and genre distributions. The comparison includes two large Web corpora that we created using a custom pipeline featuring improvements in language detection for mixed-language documents as well as boilerplate removal. Of these corpora, one is targeted at general web content (**web-dmoz**) and the other targeted more at news-style content (**web-news**). As comparison to our own Web corpora, we include a portion of deWaC [BK06] as well as text from the newspaper `die tageszeitung` (TüPP-D/Z; [Mül04]) and a recent Wikipedia dump.

Of the remainder of this paper, section 2 presents the crawling and non-linguistic preprocessing steps involved in creating the Web corpora, while section 3 presents the linguistic pipeline we used on all corpora. Our pipeline is targeted at a useful balance between processing speed, enabling its use on Web-scale corpora, and usable quality of the linguistic annotations. Section 4 describes the evaluation task chosen and presents the results of our evaluation.

2. A WEB CORPUS PIPELINE

Gathering a Web corpus consists in multiple steps: In the first phase, crawling creates an archive of HTML pages that are to be processed further; the second phase consists of boilerplate removal and deduplication, which yields the raw text of these HTML pages without any navigation elements or non-informative text; in the subsequent phases, the raw text is tokenized and sentence splitting and subsequent linguistic processing is applied.

2.1 Crawling

We use the Internet Archive's Heritrix crawler, which starts from a set of *seeds* and subsequently visits linked pages that are within the *scope* of the crawl. In deciding which page to fetch next, Heritrix maintains per-host queues that prevent any single web host from being overloaded by the crawler.

The original Web-as-Corpus approach [Sha06] uses search engine queries for sets of mid-frequent terms to gather the corresponding results as seeds. This approach crucially depends on search engine results offering a suitable quality and diversity of sources. In order to assess the role of seed selection, we created two corpora using the following seed generation strategies:

- For the **web-dmoz** corpus, we use all links from the German section of the OpenDirectory project (dmoz.org) – altogether 233 884 URLs – as seeds, and limit the crawl scope to all .de/.at domains.
- For the **web-news** corpus, we used keyword queries (for medium-frequency German words) to the Google news RSS API in order to discover appropriate seeds from relevant news sources.

  The scope of the crawl is limited to the domains of the seeds, which are usually newspapers or content-focused blog sites. Because of the seed selection strategy, the scope does include sites outside exclusively German-speaking countries (such as the German version of RIA Novosty, a Russian news agency, but also the Swiss *Bieler Tageblatt* or the Namibian *Allgemeine Zeitung*).

  In our case, we use 851 seed words which are medium-frequency nouns extracted from a newspaper corpus, and filter the results returned by RSS queries so that we have at maximum ten URLs per site in our seed list, yielding 7129 seed URLs altogether.

After crawling, all HTML pages with a file size between 4 kB and 200 kB are collected into one Zip archive per site (eliminating all identical duplicates of a given page within a site).

[1] David Graff, Christopher Cieri (2003): English Gigaword; LDC Catalogue number LDC2003T05

This process yields 48GB of compressed data (corresponding to 228 GB uncompressed) in case of the *web-dmoz* crawl, and 124 GB (540 GB uncompressed) in case of the *web-news* crawl, see also table 2.

2.2 Language and Charset Detection

For the material going into classical single-source corpora, it is very simple to make sure that the textual content is from exactly one language, encoded in a uniform fashion. In contrast, textual material from the World Wide Web exists in multiple languages, and – at least for non-English material – is represented in a variety of encodings.

The top-level domain of a web site gives a first approximation of what the language of its document may be, but is usually not very reliable: The top-level domain often correlates with its country of origin, and hence the most likely preferred language of its creator, but needs not be predictive for the language of the documents on that site. Foreign-language and mixed-language sites (even mixed-language documents, such as multilingual forums, or pages where text in one language is discussed in another) are relatively frequent. In addition, the information provided in headers by an HTTP server or declared in a page meta tags is often wrong and hence unreliable as a source of information about page encoding and language.

For accurate information, character encoding detection and language detection based on page content are required. Our approach for character encoding detection relies on the heuristics of IBM's International Components for Unicode (ICU) library,[2] while we use a customized approach for language detection that yields more adequate results for mixed-language documents.

The ICU encoding detector combines heuristics such as checking for illegal sequences for multi-byte encodings, and statistics including byte n-gram frequencies for one-byte encodings for different languages. Pages with character encoding identified with low probability (below 55%) are discarded.

Detecting the language(s) of the content is rendered more difficult by mixed-language documents as well as the presence of other-language boilerplate or nontextual data. As a result, both the exact type of heuristic used (character-level or word-level, as detailed below) and the relative ordering of language and boilerplate detection plays a role.

Two kinds of heuristics are used for language detection in general: One is using character-level properties by building a frequency profile of character n-grams and comparing this frequency profile to a reference distribution [CT94]. The other uses word-level properties of a text by using a list of common function words (which should occur in any linguistically interesting text) and makes assumptions regarding their distribution. In the latter category, [BK06] require the extracted text to contain at least 25% of function words, as well as having at least 30 tokens, and 10 different types of function words from a pre-determined list with 140 entries.

While the function word approach gives high precision in general, we found that the recall of language filtering does suffer and the resulting collection might shrink too drastically. In particular, pages that contain small amounts of useful text tend to be thrown out completely even if they constitute valid linguistic material. Because it uses global word statistics, word-based language detection is also sensitive to different-language boilerplate and must be applied after boilerplate parts of the page have been removed.

The character n-gram approach is more robust in terms of language genres, styles and domain variations, presence of the boilerplate, and size of the content. It is also likely to recognize at least one of the languages in mixed language pages, which makes it more suitable in our eyes than the word-based approach.

In our processing pipeline, we first apply a relatively loose filter in order to eliminate all the documents that are clearly non-German, keeping only German-language and mixed-language documents. The loose filter is based on character trigrams, as described by [CT94]. This first language filter is applied right after the character encoding detection. It filters out 48% of the pages in case *web-dmoz*, and 6% of the pages for the *web-news* crawl. Pages that were filtered out either contain no textual content at all, have a character encoding that could not be recognized, or were clearly identified as being from a different language.

The second step of language filtering happens after boilerplate removal and can use more precise information. In this filtering step, we apply character-level language detection to each individual block of text (corresponding roughly to one HTML paragraph). For small blocks (of less than 60 characters) and for the blocks with low language similarity, we additionally use functional word counts and the language detected for neighboring blocks in our classification.

In the case of our Web corpora, our language detection finds a significant number of mixed-language documents (numbers from *web-dmoz*):

- 80% of pages contain only German blocks.
- 11% of pages contain almost exclusively German (more than 90 % of all blocks are identified as German) .
- 8% of pages contain mostly German (more than 50% but less than 90% German content).
- 1% of pages contain mostly non-German text (less than 50% is recognized as German despite having been classified as German by the first filter language detector). These documents contain about 36% of German, 47% of English and 17% of other-language content.

For the pages containing more than 50% of German content, only the in-language blocks are kept and the remaining text is discarded. As a side-effect, much non-linguistic content, such as URLs, addresses, long list of names, math expressions, etc., which occurs in its own block, also gets discarded. The pages containing less than 50% of German content are removed completely.

[2] http://userguide.icu-project.org/conversion/detection

2.3 Duplicate and Boilerplate Detection

In order to use Web pages as a source of linguistic content, it is necessary to detect, and filter out, boilerplate text – navigation or decorative elements, advertisements, copyright notes, etc. Boilerplate text snippets would hinder both subsequent analysis steps (such as language detection) as well as distort the statistics (including distributional similarity information) of the final corpus. For similar reason, duplicated text – be it teasers for other articles, extractive summaries, or self-plagiarized content on a web page – should be removed from a Web corpus.

The most popular technique for doing boilerplate elimination, as used in the deWaC/ukWaC corpora [BBFZ09] is to find the tag node with the best ratio of (word) token and HTML tag density, and select all other segments as boilerplate. The authors point out that the approach does not handle appropriately a boilerplate in the middle of the informative content, and also can have problems at the span margins.

More refined methods include supervised learning on structural HTML features of shallow features describing a segment [KFN10], but also include computationally expensive methods such as modeling vision/position-based information [CYWM03] or wrapper induction to reconstruct formatting templates based on frequently used patterns [VdSP^{+}06].

In our case, we wanted to be able to process a wide range of genres (including, for example, forums, or user comments on other pages). We also wanted to make as few assumptions as possible on the kind of templating system used for particular pages or sites (or, similarly, absence of a templating system). To reach this goal, we implemented a content-sensitive approach to boilerplate removal that (unlike [VdSP^{+}06]) only makes very basic assumptions on the formatting and document structure.

Our approach relies on the idea that boilerplate (which may or may not be recognizable as part of the navigation based on its HTML markup) is very likely to consist in textual templates (**site patterns**). In the actual pages we would find, the site patterns are either repeated verbatim, or with several gaps filled by content that varies from pattern mention to the next (e.g., dates and user names).

Since the most reliable way to detect such site patterns is to look for repetitions in candidate patterns, there is some amount of interaction between boilerplate removal and identification of partial duplicates: Boilerplate removal may be confused by duplicate content, whereas in turn, duplicate content identification may be confused by remaining boilerplate. To mitigate these interaction problems, we interleave boilerplate filtering and near-duplicate detection: The first step uses a content-insensitive boilerplate filter, which is based on link density. In the second step, we perform the detection of near-duplicate pages within the site. The last step relies on the induction of site-specific boilerplate patterns for the content-sensitive detection of boilerplate text.

In the first step (link density filter), the number of tokens within an <a> tag is divided by the total number of tokens in the block. Our link density filter removes blocks with link density values above 0.33, which eliminates most navigational elements while leaving uncertain cases for later (content-sensitive) examination.

	link & text density			*l.d. + site patterns*			err.red.
	P	R	F1	P	R	F1	
Doc. 1	0.60	0.50	0.55	0.97	0.81	0.88	73%
Doc. 2	0.87	1.00	0.93	0.98	0.98	0.98	71%
Doc. 3	0.89	0.79	0.84	0.99	1.00	0.99	93%
Doc. 4	0.84	0.97	0.90	0.97	0.89	0.93	29%
Doc. 5	0.44	0.91	0.59	0.87	0.98	0.92	80%

Table 1: Content-sensitive boilerplate removal: Evaluation on forum HTML content

For the final step of **finding site patterns** and removing the corresponding boilerplate, we found that templates can be both as simple as blocks that repeat with an exact sequence (continuous repeats), or as complicated as blocks repeating with a sequence containing one or several gaps (discontinuous repeats). The gaps can be filled with variable length sequences that vary from pattern mention to mention (e.g. dates and nicknames, or strings such as "*On xxx yyy, zzz posted:*"). Some patterns can look like natural language text, can be lengthy, and may not be formatted specifically, thus presenting difficulty to shallow and purely structure-based approaches.

From each individual web site[3] we extract both continuous and discontinuous patterns using the GAAL library [Kis11], which performs pattern extraction using linearised suffix trees. We limit the extracted patterns to a size of not more than 500 characters, with a maximum number of three gaps, with the gaps not exceeding nine tokens in length. We consider all patterns with frequency above one to be boilerplate if the number of tokens in the repeat part(s) of a pattern exceeds the number of tokens in the gap part(s) of the pattern.

We validated our approach to content-sensitive boilerplate detection using five samples of forum HTML from different sites, which contains a mixture of quoting, signatures, and template strings intermingled with the content. Comparing boilerpipe's content-insensitive approch with our own approach (cf. Table 1), we found that using the site patterns provided substantial quality improvements.

Duplicated content (whether boilerplate or reused informative content), if present, will confuse the duplicated pattern detection. To avoid this, a **detection step for near-duplicates** (within each site) is run after the generic (link density-based) boilerplate reduction, but before the induction of site-specific boilerplate. This within-site deduplication step uses an approach identical to the near-duplicate elimination that is run on the whole corpus to find content duplicated across sites.

Deduplication uses the general approach of shingling described by [BCFM00], in that it computes a min-hash sum-

[3] A web site corresponds to one particular second-level domain, or a third-level domain inside a generic second-level one such as `.ac.at`.

	web-dmoz	web-news
raw crawl (compressed)	298GB	124GB
zip files (compressed)	48GB	112GB
zip files (uncompressed)	228GB	540GB
document counts		
before language detection	8.2M	11.4M
after language detection	3.9M	10.7M
near-duplicate removal	1.0M	5.3M
boilerplate removal	845k	4.4M
cross-site deduplication	602k	3.6M
Corpus size (tokens)	360M	1700M

Table 2: Corpus sizes after each filtering step

mary of the shingles (commonly: n-grams, for some n of 5 or 6) and uses these summaries to approximate the weighted Jaccard similarity measure between those shingles.

For our corpus pipeline, we do not use plain 5- or 6-grams as shingles, instead using an approach that creates a smaller number of still-distinctive shingles. Shingles are created based on “spots” that are composed of a function word (using articles, auxiliaries, and modal verbs in our list) and a trigram of the following three content words, as in the SpotSigs approach [TSP08]. This shingle-based representation together with shingle counts is likely to be an adequate representation for the informative content of the page.

From the shingle representation of each page the fixed length min-hash signature of the page is computed – in our case we use a min-hash signature of length 192. Min-hashing preserves the expected similarity: It has the useful property that the probability of a match at any signature index between two signatures corresponds to the Jaccard similarity of the signatures.

This property of min-hashing allows the application of locality sensitive hashing (LSH) as a next step to efficiently identify candidate near-duplicate pairs. Min-hash signatures are split into a number of bands (in our case 32) and for each band the corresponding segment (in our case segment length is 192/32=6) of the min-hash signatures is hashed to the large number of buckets. We then consider any pair that hashed to the same bucket for any of the hashings (bands) to be a candidate pair.

The SpotSigs library [TSP08] does not use straight min-hashes for a page, but uses a combination of min-hash (which represents several shingles in one hash value and is suitable for detecting exact repetition) and locality sensitive hashing (LSH), which computes a low-dimensional approximation of the content that is suitable for approximate similarity computation.

In our case, a sequence of $k = 6$ shingles (spots) is summarized into one min-hash signature, and the signatures of one document are reduced to a representation containing $l = 32$ buckets.

Locality sensitive hashing is an efficient way to approximate the similarity of the spot signatures of documents; for the actual deduplication of candidates, the weighted Jaccard similarity is calculated and any pair of documents with a similarity score of at least 44% is treated as a near-duplicate.

In the case of within-site deduplication, we reduce each cluster of mutual near-duplicates to one page that is left in the corpus and discard the other near-duplicates. For the duplicate removal across the whole corpus, we completely remove any content that has near-duplicates coming from more than five distinct sites, effectively removing this kind of frequent content. This is motivated by the observation that pages that repeat too frequently across the web present no or little linguistic interest, and frequently consist of contact information, statements of ownership and authorship of a certain text, etc.

Even using the spot-signatures approach for representing document content, boilerplate in a document representation could distort the hashed representation because of the added material. In such a case, duplicates from different sites are not recognized as such, or the boilerplate results in pages falsely recognized as duplicates. Hence, our approach of interleaving boilerplate detection (with a markup-driven first step and a content-sensitive second step) with deduplication is necessary to reach the best results.

In our corpora harvested from the Web, we find that 81% of pages are discarded with 719212 pages left in case of DMOZ crawl, and 64% of pages are discarded with 3865269 pages left in case of the news focused crawl.

3. LINGUISTIC PREPROCESSING

After extracting the raw text from a Web crawl, we can apply general-purpose NLP tools in order to extract linguistic information; In the case of Web corpora, the larger variation of genre and domain makes it necessary to use tools that are more robust than when processing only newspaper text. It is also necessary for the approaches to be reasonably efficient in order to process very large corpora.

3.1 Tokenization

From our initial evaluation of existing Web corpora, mistokenized words as well as general encoding problems were one of the concerns that we thought would most impede the final corpus quality: A mis-encoded or incorrectly tokenized word usually poses more difficulty to tagging or parsing models, especially where those models use word statistics.

Because tokenizers are not generally considered important within NLP research, it is very hard to find a common evaluation; furthermore, the models distributed with libraries such as OpenNLP exhibit rather poor performance not only because they do nothing to capture language-specific idiosyncracies but also because they are trained on inadequate data (detokenized treebank data instead of normal text).

In the case of German (where we can use data from the TüBa-D/Z treebank together with the raw text from the *tageszeitung* newspaper), tokenization exhibits nonlocal properties in the case of compounds such as ‘*”Sicherheits”-Truppen*’ or ‘*(Schaden-)Freude*’. In examples such as these, quotes or parentheses have to be kept together when they are part of a compound, but not when they connect normal text. OpenNLP’s approach based on local classifiers would always introduce errors in such case, whereas a rule-based tokenizer can reasonably be expected to solve such a problem.

If one wanted to reach perfect agreement with the treebank tokenization, additional semantic knowledge would be necessary to distinguish coordination-like use of dashes (‘*5-10 Gramm*’), which has to be represented as separate tokens, with dashes as token-internal separator in room and telephone numbers (‘*Raum 5-10*’).

Especially for multi-source data, but also for normal newspaper text, it has to be kept in mind that Unicode contains a multitude of alternatives for various nonalphabetic characters, especially dashes and quotes. Common processing tools such as part-of-speech taggers and parsers are best used with text that is similar to the treebank data they have been trained on. In particular, treebank text normally contains only normal quotes and dashes instead of reproducing all and any typographic variation. In order to provide reasonable input to the rest of the processing chain, we took a very pragmatic approach and normalized all quotes and dashes to their standard ASCII forms. We find that the additional typographic information is not important enough to accept (or deal with) the resultant fallout in the rest of the linguistic processing pipeline.

In order to provide the tokenization for our Web corpora, we used a rule-based tokenizer, which first performs an initial segmentation of a text into tokens using a set of regular expressions. In a subsequent step, the tokenizer revises some of these tokenization decisions and also performs sentence boundary detection.

While Unicode-compatible, the regular expressions library included with the standard Java library is not very performant since it needs to support the backtracking needed for a Perl-compatible treatment of regular expressions; as a result, very complicated regular expressions (including cases typical for tokenization, such as lists of abbreviations, or nesting of alternatives for different sub-parts of a token) show some performance deterioration. The `dk.brics.automaton` library[4] allows to use deterministic finite automaton (DFA) representation for matching of expressions in addition to supporting Unicode and the corresponding character classes.

3.2 Morphology and Parsing

For fixed word order languages such as English, approaches to induce semantic information from raw text have been shown to be feasible, even if they do not always reach the performance of more elaborate approaches. In general, however, deeper linguistic information including lemmatization and phrase structure or dependency parses can be expected to yield much better generalization behaviour. This is especially true for languages such as German, which combine a more flexible word order with rich morphological inflection behavior.

For English and its fixed word order, [CM02] have proposed the use of a chunker based on sequence tagging and subsequent chunk linking in order to gain information on grammatical relations that can be used as describing context for nouns, in particular premodifying adjectives, subjects, and direct objects. In the case of German, syntactic properties of the language make a chunking-based approach problematic, more so as a simple chunk linker would be unable to recover most argument relations (subject, object) in a language with somewhat free word order.

As a fast compromise between (insufficiently powerful) chunking and (slow) full parsing, we use a pipeline that relies on deterministic dependency parsing to provide complete dependency parses at a speed that is suitable for the processing of Web-scale corpora.

The parsing model is based on MALTParser, a transition-based parser, and uses part-of-speech and morphological information as input. Morphological information is annotated using RFTagger [SL08], a state-of-the-art morphological tagger based on decision trees and a large context window (which allows it to model agreement more accurately than a normal trigram-based sequence tagger). While transition-based parsers are quite fast in general, an SVM classifier (which is used in MALTParser by default) becomes slower with increasing training set. In contrast, using the MALTParser interface to LibLinear by [Cas09], we can reach a much larger speed of 55 sentences per second (against 0.4 sentences per second for a more feature-rich SVM-based model that reaches state of the art performance).

For lemmatization, the original version of deWaC uses the lemmatization component of TreeTagger [Sch95], which is easily the most popular lemmatization component for German since it is freely available (at least for noncommercial purposes) and easy to use. Several weaknesses of TreeTagger make it less than ideal for the use in a pipeline for learning lexical semantic information: firstly, TreeTagger uses a fixed lexicon and provides no treatment for unknown words; secondly, it provides no solution for reattaching separable verb prefixes, yielding only partial verb lemmas in many cases; thirdly, it always provides the same lemma for one word/POS combination and does not use morphosyntactical information for disambiguating lemmas.

In our case, we use the syntax-based TüBa-D/Z lemmatizer [VBHT10], which uses a separate morphological analyzer and some fallback heuristics. The compositional and derivational SMOR morphology [SFH04] serves to provide morphological analyses for novel words. For unanalyzed novel words that are not covered by SMOR, the lemmatizer falls back to surface-based guessing heuristics. It uses morphological and syntactic information to provide more accurate lemmas; In addition to dependency structures, the morphological tags from RFTagger as well as global frequency information are used.

4. EVALUATION

For English newspaper text, [CM02] provide an account of the relation between corpus size and processing algorithms used on one hand, and the quality of the learned thesaurus as well as the amount of computation that is necessary for the preprocessing on the other hand.

Because it provides a unified framework for evaluating a combination of the (kind and amount of) text used as input,

[4] http://www.brics.dk/automaton/

as well as the processing tools used, Curran and Moens' experimental framework provides an excellent way to compare different ways of acquiring text (such as a fixed amount of newspaper text versus a much larger amount from a Web corpus). It also allows to explore the most efficient way to deal with resource constraints: In the case of a fixed amount of computing power being available, they find that window-based approaches, despite being computationally very cheap, yield results that are significantly inferior to applying more elaborate techniques on a smaller amount of text. In comparison, their approximate chunk linking technique works well enough for English that it may be preferred to a full parser when computation time is limited.

4.1 Single-source Corpora

Besides Google's n-gram dataset for German, which necessitates a different approach from the corpora we use here (as it is only usable for shallow pattern extraction), we use a 600 million word subset from the deWaC corpus of [BK06] and two single-source corpora that would be used alternatively to a Web corpus: on one hand, about 200 million words of newspaper text extracted from the years 1986-1999 of *die tageszeitung*; on the other hand, a recent Wikipedia dump (amounting to about 400 million words) that has been cleaned using the Tanl Wikipedia extractor[5] before applying our linguistic processing pipeline.

4.2 A German 'Gold Standard' Thesaurus

[CM02] use several existing thesauri for English, namely the Macquarie, Roget's and Moby thesauri, to create a merged thesaurus that provides a gold standard for "matching synonyms" that a system should retrieve. To our best knowledge, such thesauri do not exist for German (and may be nonexistent or hard to get for other languages), which makes it necessary to adapt their approach.

Using GermaNet 6.0 [HH10], we found that looking for exact synonyms (i.e., words that occur in one of the synsets of the target word) yields very narrow synonym lists, which are often focused on non-dominant senses of a word. Thesauri, in contrast, contain broader lists of related terms and also include near-synonyms (i.e., cohyponyms). In order to get lists of related terms that are better suited for evaluating semantic similarity measures than just using synonyms, we adapted the method of *radial glosses* [Gur05] for our purpose.

Gurevych discusses a number of possibilities to extract super- and/or subordinate and coordinate terms from a wordnet; for our purpose, we aimed at words of roughly the same level of generality and sharing an appropriately high number of properties. To implement this, we ordered neighbouring synsets using a path distance measure and added the words of the closest unused synset until 30 words have been retrieved. For the distance measure itself, we postulate higher distances for links between very general terms (especially the top three layers of GermaNet's noun hierarchy), and for links that have a large number of sister links (such as those that link a particular kind of animal to the individual species that make up this kind).

To yield a testing set for the evaluation of corpora and processing pipelines, we took samples of target words consisting of the top 30 items (by frequency in the TüPP-D/Z corpus), and additional 30 items from the frequency ranges of 10-20, 5-10, 2-3, 1-2 occurrences per million. (For a 200 million word corpus, these would correspond to around one million occurrences among the most frequent words, and between 3000 and 300 occurrences of the sampled words with lower frequency of occurrence).

4.3 Distributional Thesaurus Construction

In order to compare corpus quality via the intermediary of distributional semantics tasks, we create vectors of co-occurring lemmas based on various methods. The vector entry for each co-occurring lemma is weighted based on the conservative pointwise mutual information estimate of [PR04]. To retrieve semantically similar words based on the vector representation, we then use a similarity measure based on the Jensen-Shannon divergence to compute a ranked list of (frequent) distributionally similar terms. The Jensen-Shannon divergence was found by [PKS04] to yield the best results among several alternative vector similarity functions.

For computing the vectors, we extracted related lemmas using two methods: The first method extracts weighted collocations in the way described by **Padó and Lapata's method** [PL07], which weights (same-clause) collocates by their distance in the dependency graph (yielding 1, $\frac{1}{2}$, or $\frac{1}{3}$ for nodes that are 1, 2, 3 edges away, respectively).

The second method uses selected **grammatical relations** extracted from the dependency parses and uses the frequency profile of terms co-occurring with one specific relation to construct the word vector. We found that premodifying adjectives (the `ATTR` label in the dependency parses) performed best among all relations, with accusative objects (`OBJA`) also doing fairly well.

As a means to make use of Google's German n-gram data, we used **shallow patterns** consisting of a sequence of one known adjective and one known noun, or a coordination of two known nouns, using a simple precomputed lemma mapping.

4.4 Evaluation Results

Using a distributional similarity evaluation gives us a way to compare the effect of varying the text basis as well as the methods for computing the actual distributional similarity values. Looking at the results from table 3, we see that our evaluation results are in fact better for the `tageszeitung` corpus (TüPP-D/Z), despite avoiding any kind of explicit bias. [6]

Table 6 compares different methods for collocate extraction both on dependencies from TüPP-D/Z corpus as well as pat-

[5] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

[6] Implicit sources of bias may be seen in the frequency lists used for choosing words for addition to GermaNet, which have been constructed from varied sources including *tageszeitung* and the German Wikipedia, but also by the fact that the frequency-based sampling of words was performed based on frequency data numbers from the `tageszeitung` corpus.

corpus	200m	400m	600m
TüPP-D/Z	0.69	—	—
Wikipedia	0.67	0.68	—
web-dmoz	0.66	0.66	—
web-news	0.64	0.66	0.66
deWaC	0.68	0.70	0.69

Table 3: InvR scores for different subcorpus sizes, Padó and Lapata's method

corpus	200m	400m	600m
TüPP-D/Z	0.88	—	—
Wikipedia	0.77	0.84	—
web-dmoz	0.77	0.84	—
web-news	0.80	0.86	0.89
deWaC	0.82	0.87	0.90

Table 4: InvR scores for different subcorpus sizes, premodifying adjectives

corpus	200m	400m	600m
TüPP-D/Z	0.71	—	—
Wikipedia	0.62	0.72	—
web-dmoz	0.58	0.64	—
web-news	0.63	0.72	0.75
deWaC	0.61	0.71	0.74

Table 5: InvR scores for different subcorpus sizes, accusative objects

terns extracted from Google's German n-gram dataset. It is quite surprising that the numbers extracted from the larger n-gram dataset are substantially lower, since the n-gram corpus was constructed using about 500 times more data (100 billion words versus 200 million words). It is also notable that the grammatical relation approach is decidedly superior to the one using simple (dependency-)syntactic neighbourhood collocates.

Looking at the lowest frequency range shows that Padó and Lapata's method of collecting collocates independent of a particular location is much less sensitive to sparse data problems – indeed the results do not seem to improve with corpus size – while the grammatical relation approach as well as the shallow n-gram patterns decidedly show deficiencies in the lowest frequency range. (In the frequency range from 1-2 occurrences per million tokens, Padó and Lapata's approach is the strongest feature extraction method for all corpora except for the newspaper corpus, where it ranks behind premodifying adjectives).

The grammatical relation approach indeed profits from larger corpus size, leading to improvements even over the newspaper corpus in the case of the larger Web corpora. One important question would be if the frequency selection in Curran and Moens' work (which we used as a starting point for our own gold standard) is really typical for real-world applications in that it contains very few low-frequency terms, hence being less sensitive to data sparseness than other tasks would be.

method	corpus	InvR
PL07	TüPP-D/Z	0.69
Adj-N	web-news	0.89
Adj-N	ngrams	0.57
N-and-N	ngrams	0.51

Table 6: Parsed corpora vs. n-gram patterns

5. SUMMARY

In this work, we have extended the evaluation framework for English corpora and processing used by [CM02] for use with German Web corpora and have used it to compare different approaches to building corpora and extracting features for distributional semantic models. We have used this evaluation framework to provide evidence for the quality of our pipeline for building textual corpora from samples of the World Wide Web. The pipeline incorporates several improvements on the state of the art. In particular, we propose a novel approach to boilerplate removal which uses the crawled data for one site to realize content-sensitive filtering of boilerplate-heavy text, and present state-of-the-art techniques for linguistic processing of the resulting text.

The results from our evaluation procedure show that producing and using Web corpora in general yields better results than using shallow pattern search on Google's German n-gram dataset, and further that different methods for the construction of vector representations show considerable difference in their sensitivity to the amount of corpus data available.

For researchers aiming to create Web corpora in other languages than German, a number of insights can be formulated: One is that, even in the presence of Google's n-gram collections for many languages, a Web corpus affords more complex linguistic processing and may be considerably more useful for many purposes.

Secondly, it is apparent that methods such as that of Padó and Lapata are at a disadvantage to the explicit modeling of grammatical relations as soon as there is sufficient data for the latter to work; in our evaluation, as well as that of Curran and Moens [CM02], the sampling of test words is biased towards higher-frequency items and may overemphasize this tendency even at moderate corpus sizes. Lastly, the processing pipeline based on deterministic dependency parsing with a linear classifier (together with a step of morphological analysis) that we use here scales well to Web-scale corpora and may present an appropriate choice for many languages.

Acknowledgements The research reported in this paper was partially funded by the Deutsche Forschungsgemeinschaft as part of collaborative research centre (SFB) 833. The authors would like to thank Emily Jamison for helpful comments on an earlier version of this paper.

6. REFERENCES

[BBFZ09] Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The Wacky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009.

[BCFM00] Andrei Z. Broder, Moses Charikar, Alan M Frieze, and Michael Mitzenmacher. Min-wise independent permutations. *Journal of Computer and System Sciences*, 60(3):630 – 659, 2000.

[BK06] Marco Baroni and Adam Kilgariff. Large linguistically-processed web corpora for multiple languages. In *EACL 2006*, 2006.

[BL08] Marco Baroni and Alessandro Lenci. Concepts and word spaces. *Italian Journal of Linguistics*, 20(1):129–156, 2008.

[Cas09] Sofia Cassel. MaltParser and LIBLINEAR - transition-based dependency parsing with linear classification for feature model optimization. Master's thesis, Uppsala University, 2009.

[CM02] James Curran and Marc Moens. Scaling context space. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.

[CT94] William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.

[CYWM03] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. Extracting content structure for web pages based on visual representation. In *Proc. 5th Asia Pacific Web Conference (APWeb)*, volume 2642 of *LNCS*, 2003.

[FP10] Manaal Faruqui and Sebastian Padó. Training and evaluating a German named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*, 2010.

[Gur05] Iryna Gurevych. Using the structure of a conceptual network in computing semantic relatedness. In *IJCNLP 2005*, 2005.

[HH10] Verena Henrich and Erhard Hinrichs. GernEdiT - the GermaNet editing tool. In *LREC 2010*, 2010.

[KFN10] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, 2010.

[Kis11] Alexander Kislev. Automated multilingual alignment of discontinuous sequences. Master's thesis, Seminar für Sprachwissenschaft, Universität Tübingen, 2011. See also http://code.google.com/p/gaal/.

[LC06] Vinci Liu and James R. Curran. Web text corpus for natural language processing. In *EACL 2006*, 2006.

[Mül04] Frank Henrik Müller. Stylebook for the Tübingen partially parsed corpus of written German (TüPP-D/Z). Technischer Bericht, Seminar für Sprachwissenschaft, Universität Tübingen, 2004.

[PKS04] Viktor Pekar, Michael Krkoska, and Steffen Staab. Feature weighting for co-occurrence-based classification of words. In *COLING'2004*, 2004.

[PL07] Sebastian Padó and Mirella Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007.

[PR04] Patrick Pantel and Deepak Ravichandran. Automatically labeling semantic classes. In *HLT/NAACL 2004*, 2004.

[Sch95] Helmut Schmid. Improvements in part-of-speech tagging with an application to German. In *Proc. ACL-SIGDAT Workshop*, 1995.

[SFH04] Helmut Schmid, Arne Fitschen, and Ulrich Heid. SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of LREC 2004*, 2004.

[Sha06] Serge Sharoff. Open-source corpora: using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4):435–462, 2006.

[SL08] Helmut Schmid and Florian Laws. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *COLING 2008*, 2008.

[TSP08] Martin Theobald, Jonathan Siddharth, and Andreas Paepcke. SpotSigs: robust and efficient near duplicate detection in large Web collections. In *Proc. 31st SIGIR conference on Research and development in information retrieval*, pages 563–570. ACM, 2008.

[VBHT10] Yannick Versley, A. Kathrin Beck, Erhard Hinrichs, and Heike Telljohann. A syntax-first approach to high-quality morphological analysis and lemma disambiguation for the TüBa-D/Z treebank. In *Proceedings of the 9th Conference on Treebanks and Linguistic Theories (TLT9)*, 2010.

[VdSP$^+$06] Karane Vieira, Altigran A. da Silva, Nick Pinto, Edleno S. de Moura, Ao M. Jo, and Juliana Freire. A fast and robust method for web page template detection and removal. In *Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM'06)*, 2006.

[Ver08] Yannick Versley. Decorrelation and shallow semantic patterns for distributional clustering of nouns and verbs. In *ESSLLI 2008 Workshop on Distributional Lexical Semantics*, 2008.