Proceedings of the 8th Web as Corpus Workshop (WAC-8) @Corpus Linguistics 2013

Stefan Evert[1], Egon Stemle[2] and Paul Rayson[3]

[1] Friedrich-Alexander-Universitt Erlangen-Nrnberg (FAU), Erlangen, Germany
[2] European Academy of Bozen/Bolzano (EURAC), Bolzano (BZ), Italy
[3] Lancaster University, Lancaster, U.K.

Web corpora and other Web-derived data have become a gold mine for corpus linguistics and natural language processing. The Web is an easy source of unprecedented amounts of linguistic data from a broad range of registers and text types. However, a collection of Web pages is not immediately suitable for exploration in the same way a traditional corpus is.

Since the first Web as Corpus Workshop organised at the Corpus Linguistics 2005 Conference, a highly successful series of yearly Web as Corpus workshops provides a venue for interested researchers to meet, share ideas and discuss the problems and possibilities of compiling and using Web corpora. After a stronger focus on application-oriented natural language processing and Web technology in recent years with workshops taking place at NAACL-HLT 2010, 2011 and WWW 2012 the 8th Web as Corpus Workshop returns to its roots in the corpus linguistics community.

Accordingly, the leading theme of this workshop is the application of Web data in language research, including linguistic evaluation of Web-derived corpora as well as strategies and tools for high-quality automatic annotation of Web text. The workshop brings together presentations on all aspects of building, using and evaluating Web corpora, with a particular focus on the following topics:

- applications of Web corpora and other Web-derived data sets for language research
- automatic linguistic annotation of Web data such as tokenisation, part-of-speech tagging, lemmatisation and semantic tagging
- (the accuracy of currently available off-the-shelf tools is still unsatisfactory for many types of Web data)
- critical exploration of the characteristics of Web data from a linguistic perspective and its applicability to language research
- presentation of Web corpus collection projects or software tools required for some part of this process (crawling, filtering, de-duplication, language identification, indexing, ...)

Table of Contents

Feed Corpus : An Ever Growing Up-To-Date Corpus

Akshay Minocha
IIIT Hyderabad
India
akshay.minocha@students.iiit.ac.in

Siva Reddy
Lexical Computing Ltd
United Kingdom
siva@sketchengine.co.uk

Adam Kilgarriff
Lexical Computing Ltd
United Kingdom
adam@lexmasterclass.com

Abstract

Corpus-based lexicography has to keep its pace with the language evolution by using up-to-date corpus. In this paper we propose a method for collecting corpora which is ever growing and up-to-date with the language. We make use of social media to discover sources of dynamic content like blogs and news websites. Most such websites provide a short summary of their content change in a separate page known as feed, which we use to keep track of new content. We collect millions of such feeds using social media. We design a scheduler which rank these feeds based on their frequency of update and the amount of text extracted per retrieval. Based on the rank, we periodically crawl these feeds and add any new content generated to our corpus collection along with the temporal information. Thus the corpus is dynamic and nearly up-to-date with the language evolution. In a month's duration, we collected a corpus of size 1.36 billion words for English, which after deduplication resulted in 300 million words, demonstrating the potentiality of this approach.

1 Introduction

The aim of corpus-based lexicography is to study the behaviour of words based on their usage in language. Since the success of COBUILD project, many static collections of electronic corpora such as BNC, the WaCky Corpora (Sharoff, 2006; Baroni et al., 2009; Kilgarriff et al., 2010), and recently TenTens' (Jakubíček et al., 2013) came into existence. The list has been increasing each year and soon overwhelming with too many choices for the publishing agencies. With the change in corpus choice, the lexicographic cycle has to be rerun, although the major part of the corpus contains the language usage already captured in the past. Instead what if the corpus is dynamic and provides a snapshot of language usage between any two desired periodic points?

In this paper, we propose a solution for creating dynamic corpora which is up-to-date with the language and increasing every day. There were past proposals on dynamic corpora like monitor corpus (Clear, 1987) which is a continuous stream of corpus rather than a static collection. With the current advances in web technology, building such corpora is not far from achievable.

Traditional methods of static corpora collection involve crawling through billions of web pages periodically. Keeping track of changes in such a huge network is pain-staking, and visiting each website again in the the next crawl anticipating for new content is cost-inefficient. Technology such as feeds partly address this problem by providing a mechanism to detect new content. A feed is a collection of temporal updates in a website. The presence of a feed in a website is a plausible indication that the website posts new content once-in-a-while. We aim to fish millions of feeds from the Internet and keep track of the changes and add new content to the corpus collection whenever a website is updated. But how to discover millions of feeds from several millions of websites?

With the advent of social media like Twitter, latest content is one click away. Currently 340 million tweets are published per day[1]. This is expected to increase tremendously given that 87% of all the tweets are posted in the past 24 months (Leetaru et al., 2013). Most newswires, blogs and other frequently updated websites post tweets whenever new content is generated. Additionally, millions of people share hyperlinks of posts containing their newly found information. These tweets are potential sources to the websites which

[1]Twitter Statistics - http://blog.twitter.com/2012/03/twitter-turns-six.html

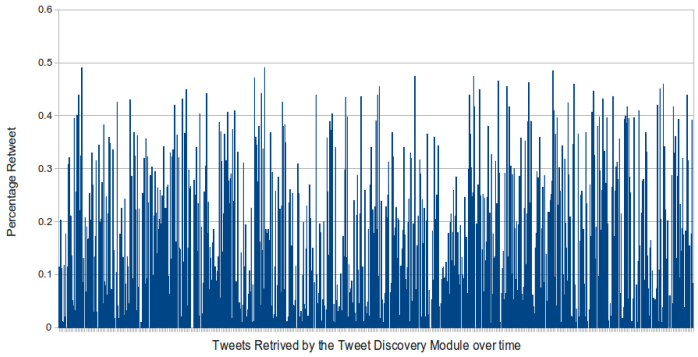


Figure 1: Number of retweets on the scale of 0-1 in the month of March 2013

contain feeds, and once a feed is found, we track it for new content updates. We piggyback Twitter by retrieving tweets containing hyperlinks using keyword search queries, thus in-turn piggybacking on the world's population to find websites that frequently update, saving ourselves from the tedious task of crawling the whole web.

To keep track of millions of feeds, a simple sequential feed aggregator is not sufficient since different feeds have different update frequencies and a long looping delay will lose novel content from frequently updated websites. We design a scheduler which sets a priority for each feed based on its frequency of update between posts and the amount of content generated per retrieval. Based on the scheduler priorities, we fetch feeds and check if any new posts are posted on the website. The new content from posts is added to the corpus collection, resulting in an corpus ever-growing and relatively up-to-date corpus compared to the static corpora.

In our pilot experiment on English run for the month of March 2013, we collected around 1.36 billion words, which after duplicate removal resulted in 300 million words. Currently, we have up to 150 thousand feeds and still growing.

2 Method for Feed Corpus Collection

We briefly outline the steps involved in our approach for collecting an ever-growing up-to-date corpus, the Feed Corpus.

1. **Feed Discovery:** For every 15 minutes, we run keyword search queries on Twitter to retrieve tweets containing hyperlinks. The keywords are chosen such that they result in tweets with hyperlinks to potential websites containing feeds.

2. **Feed Validation:** The domains of the hyperlinks in the tweets from Step 1 are validated for the presence of a feed. If a feed is found, the feed is added to the list of valid feeds or else the domain is blacklisted.

3. **Feed Scheduler:** A feed contains up to top 10 last new posts along with the timestamps at which these posts are created. An initial priority is set to the feed based on its frequency of update and placed in the priority queue.

4. **Feed Crawler:** The top most feed in the priority queue is fetched and verified for any new posts. If a new post is found and is not already in the corpus (content deduplication), the post is added to the corpora collection along with its time stamp. The priority of the feed is updated based on its previous priority and the new average time lapse between posts.

In the coming subsections, we describe each step of feed corpus collection and the corresponding challenges involved.

2.1 Feed Discovery

We design Twitter search queries containing one of the keywords like news, business, arts, games, regional, science, shopping, society etc. We restrict the search results to tweets containing hyperlinks in them. All these queries are run in periodic intervals of about 15 minutes. We ignore retweets since these likely result in duplicates. Figure 1 displays the number of retweets for our search query results on the scale of 0-1 for a duration of a month. Almost 17 percent of the query results constitute retweets.

The hyperlinks we find in the tweets are not the feed links themselves but to the posts that could belong to a domain containing feeds. We validate the feeds in the next step.

2.2 Feed Validation

Given a hyperlink of a web post, this module determines the feed links associated with the hyperlink. Hyperlinks shared on Twitter are unlikely to be feed links themselves. We use simple heuristics to determine the feed link of a hyperlink. In the first step, the hyperlink is verified if it is a feed link by itself. If not, we analyse the main domain of the hyperlink and find out references to any valid feed hyperlinks mentioned in the source. We use meta-data commonly associated with the feeds to identify feeds e.g. `type=application/rss+xml`. If the main domain has no feed links, we analyse the child hyperlink one step above the main domain along the initial hyperlink. For example, if the initial hyperlink is `http://ab.cd.com/ef/ij/kl.rss`, the *main domain* is `http://ab.cd.com/` and *one step above main* is `http://ab.cd.com/ef`.

In case if no feed links are found, we blacklist the domain to avoid repeating these steps in future if any of the hyperlinks associated with this domain are seen again.

2.3 Feed Scheduler

Different feeds have different update times. For example, newswire websites post new content every hour, while corporate websites post content once in a day or two, and personal blogs are updated once in a week or a month or even a year. After Step 2, we end up with millions of feeds over time. A simple sequential aggregator is inefficient in tracking updates: if the time gap between visits to a feed is higher than its frequency of update we lose its updates; if the time gap is lower than its frequency of update, we waste bandwidth. A scheduler is crucial to build an efficient crawler. We determine the initial update frequency of a feed by taking an average time gap between the top 10 recent posts. We implement a time-based priority queue where all the feeds are placed in the queue according to the priority. As the time passes the queue moves forward with the first feed in the queue processed by the feed crawler.

Additionally, we aim to avoid crawling websites which frequently change but results in low yield rate e.g. if the website has too much boilerplate or the language is irrelevant.

Threshold θ ($Y * 10^{-2}$)	Crawler Output size (GB)	Final data size (GB)
0.01	221.16	1.59
0.04	162.52	1.54
0.32	72.31	1.48
2.56	14.36	0.90
5.12	6.71	0.72

Table 1: Amount of data collected with different yield rates in the month of March 2013

We use the yield rate as described by Suchomel and Pomikálek (2012)

$$\texttt{yield rate Y} = \frac{\textit{cleaned data size}}{\textit{downloaded data size}}$$

The yield rate signifies the efficiency of fetching data from a website. We discard the domains whose yield rate (Y) is below a threshold (θ). We use jusText[2] (Pomikálek, 2011) for removing boilerplate text from web pages and langid.py (Lui and Baldwin, 2012) to detect language of the page.

Table 1 displays statistics of corpora downloaded with different thresholds θ of yield rates. We choose a yield rate of 0.003 to allow data even from micro-blogging sites.

2.4 Feed Crawler

In this step, we take the highest priority feed in the priority queue and crawl it to detect any new content. If found, the content is added to the corpus collection after verifying if it is a duplicate of any post already present in the data. We use the deduplication tool Onion[3] (Pomikálek, 2011) to remove duplicates. Since it is expensive to run the deduplication tool for every new post, we run the tool in a batch mode after collecting significant amount of new content.

We also update the yield rate and priority rank of the feed in the scheduler. Feeds do not have a constant update time. We observed bursts of activity occasionally and a period of long silence. In order to take past into consideration, we update priorities in a cumulative fashion taking a weighted mean of the past update time and the current update time as new update time. Based on the new update time, we place the feed in the priority queue.

[2] jusText https://code.google.com/p/justext/
[3] http://code.google.com/p/onion/

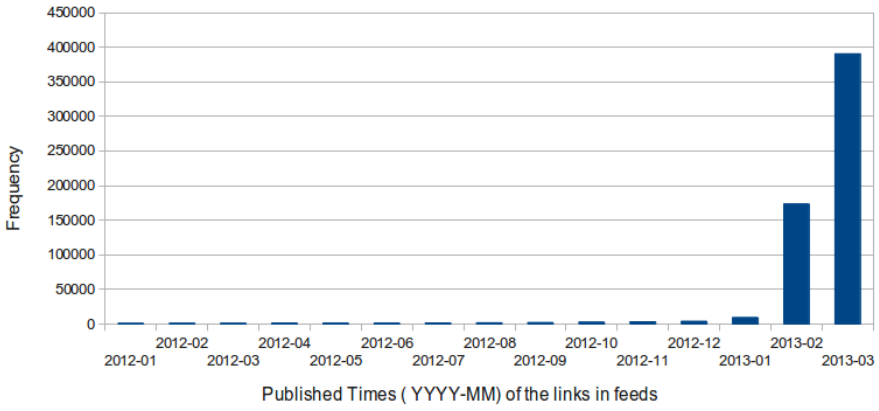


Figure 2: Number of posts from each month in our corpus

3 Results

As a pilot experiment, we ran the above steps for the month of March 2013. The results are preliminary and we hope to extend this work in future. At each iteration of feed discovery, we found around 120 new hyperlinks of which 60% helped in finding valid feeds. We collected around 150,000 feeds in a month which shows Twitter is promising to discover new feed links. On an average we collected around 40 million words per day from the feeds. We collected the corpus from all the posts listed in the feeds, including the posts from previous months of March. Figure 2 displays the number of posts collected for each month. Since we collected feeds from tweets in March, most posts are found to be from March. At the end of the month, we ran deduplication on total corpus of size 1.3 billion words and extracted cleaned corpus of size 300 million words.

4 Conclusion

We presented a simple method for building an ever-growing up-to-date corpora using feeds discovered from Twitter. In a month's duration, we collected around 150,000 feeds and a corpus of 300 million words along with their timestamps of creation. Our preliminary results are promising encouraging us to extend this work for many other languages and over a prolonged period of time. Currently, we only use the temporal information present in the feeds, but in future we also aim to use category tags mentioned in the feeds, thus providing a valuable resource for genre-specific temporal corpora.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Jeremy Clear. 1987. Trawling the language: monitor corpora. *ZURILEX Proceedings. Tübingen: Francke*.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlỳ, and Vít Suchomel. 2013. The tenten corpus family. In *International Conference on Corpus Linguistics, Lancaster*.

Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and Avinesh Pvs. 2010. A corpus factory for many languages. *Proc. LREC, Malta*.

Kalev Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. 2013. Mapping the global twitter heartbeat: The geography of twitter. *First Monday*, 18(5).

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30. Association for Computational Linguistics.

Jan Pomikálek. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora*. Ph.D. thesis, Ph. D. thesis, Masaryk University.

Serge Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. *WaCky*, pages 63–98.

Vit Suchomel and Jan Pomikálek. 2012. Efficient web crawling for large text corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 39–43.

LWAC: Longitudinal Web-as-Corpus Sampling

Stephen Wattam[1] **Paul Rayson**[1] **Damon Berridge**[2]

Lancaster University
School of Computing and Communications[1], Mathematics and Statistics[2]
f.lastname@lancs.ac.uk

1 Sampling

Many sampling efforts for linguistic data on the web are heavily focused on producing results comparable to conventional corpora. These typically take two forms: those based on URI lists (e.g. from search results, as in BE06 (Baker, 2009), BootCat (Baroni and Bernardini, 2004)), and those formed through crawling (e.g. enTenTen[1], UKWaC (Ferraresi et al., 2008)).

Though initial efforts in web-as-corpus (WaC) focused on the former method many projects are now constructing supercorpora, which may themselves be searched with greater precision than the 'raw' web, in line with Kilgarriff's vision of linguistic search engines (Kilgarriff, 2003). This has led to the proliferation of crawlers such as those used in (Schäfer and Bildhauer, 2012) and WebCorp[2].

This approach, with its base in a continually-growing supercorpus, parallels the strategy of a monitor corpus, and is applicable to linguistic inquiry concerned with diachronic properties.

Repeated sampling by crawling, whilst balanced linguistically, omits subtler technical aspects that govern consumption of data online, most notably the user's impression of its location, as defined by the URI. Low publishing costs online, paired with increasing corporate oversight and reputation management (both personal and professional), have lead to a situation where this content is being revised frequently, often without users even noticing.

The nature of within-URI change have been studied from a technical perspective by those interested in managing network infrastructure, compiling digital libraries, and optimising the maintenance of search engine databases. The needs of these parties are quite aside from those of corpus researchers, however, since they focus around a best-effort database of information, rather than a dependable longitudinal sample with known margins for error.

We present here a tool, LWAC, for this form of longitudinal sampling, designed to maximise the comparability of documents downloaded in each sample in terms of their URI rather than content. To accomplish this, we use a batch-mode sampling strategy, as illustrated in Figure 1, to get full coverage over a list of URIs, at the expense of sampling new content.

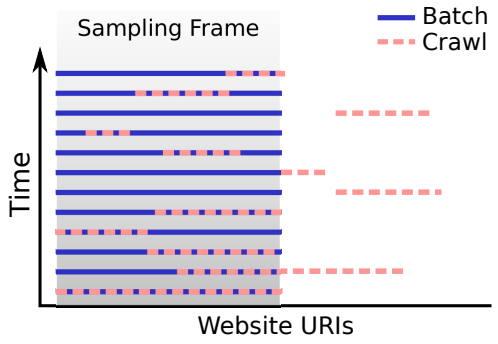


Figure 1: URI coverage for batch and crawl

2 Applications

Our strategy allows us to investigate how language may change in relation to technical and social events in a way that mimics the experience of the end user, and offers a useful perspective on many epistemic problems of WaC methods, to determine:

- The portions of web pages that typically change as main content regions;
- The impact of social feedback and user generated content on page content;
- How censorship, redaction and revision affect website contents;

[1] http://trac.sketchengine.co.uk/wiki/\Corpora/enTenTen

[2] http://www.webcorp.org.uk/live/

- Website resource persistence and its relation to linguistic content (link rot/document attrition);
- How institutions' publishing policies affect reporting of current events.

In order to maximise its coverage of these topics, LWAC is designed to construct longitudinal samples from arbitrary URI lists, using commodity hardware, in a way that mimics the user's experience of a website.

3 Architecture & Performance

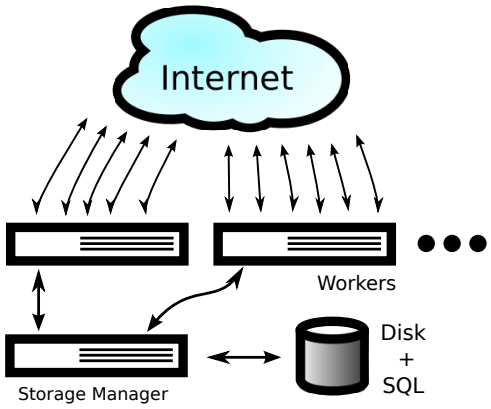


Figure 2: System Architecture

In order to form a useful longitudinal sample, each data point should be as time-invariate as possible. As such, a highly parallel, distributed architecture was selected (Figure 2). This yields technical benefits in terms of throughput (especially where the internet connection is a bottleneck), flexibility, and the ability to differentiate between websites that are blocked for a given area of the internet and those that are offline proper.

Data storage in the system is split between metadata, stored in an SQL database, and website sample data itself, which is stored as raw HTTP response data in a versioned structure on disk. The storage format is optimised for large samples, and is nested in order to avoid common filesystem limits. LWAC does not enumerate URIs in memory, meaning there is no hard limit on corpus size.

The download process is managed by a central server, which co-ordinates storage and metadata access and provides full atomicity. This server distributes batch jobs, according to policies governing reliability and throughput, to worker clients, which compete for the opportunity to download web pages.

Workers are able to imitate the behaviour of end users' browsers as much as possible, so as to avoid search engine optimisation and user-agent detection tactics (for example, they may retain cookies and present typical request headers).

After downloads have occurred, data may be retrieved for analysis in a variety of formats using the included export tool.

In practice throughput is limited by several factors, among them the available bandwidth, number of worker clients, speed of DNS lookups, and the proportion of links which are destined to time out during connection. With favourable network conditions, each worker is capable of downloading a sample the size of the BE06 corpus every four seconds. Servers can support any number of workers, but this is limited in practice by the bandwidth available to the server relative to that available for web requests.

4 Conclusion

The LWAC sampling tool, available online[3], offers an easy and rigorous way to compile longitudinal web corpora from arbitrary URI lists. We believe it has particular utility to investigation of challenges that face WaC methods, as well as fine-grained sampling of language linked to current events and other fast-moving phenomena.

References

[Baker2009] Paul Baker. 2009. The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics*, 14(3):312–337.

[Baroni and Bernardini2004] Marco Baroni and Silvia Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *LREC Proc.*, volume 4.

[Ferraresi et al.2008] Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating UKWaC, a very large web-derived corpus of English. In *The 4th Web as Corpus Workshop (WAC-4) Can we beat Google? Proc.*, pages 47–54.

[Kilgarriff2003] Adam Kilgarriff. 2003. Linguistic search engine. In *proceedings of Workshop on Shallow Processing of Large Corpora (SProLaC 2003)*, pages 53–58.

[Schäfer and Bildhauer2012] Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *LREC Proc.*, volume 8.

[3] `http://ucrel.lancs.ac.uk/LWAC/`

The Good, the Bad, and the Hazy: Design Decisions in Web Corpus Construction

Roland Schäfer
Freie Universität Berlin
roland.schaefer@fu-berlin.de

Adrien Barbaresi
Freie Universität Berlin
adrien.barbaresi@ens-lyon.fr

Felix Bildhauer
Freie Universität Berlin
felix.bildhauer@fu-berlin.de

Abstract

In this paper, we examine notions of text quality in the context of web corpus construction. Web documents often contain material which disqualifies them from inclusion in a corpus (tag clouds, lists of names or nouns, etc.). First, we look at the agreement between coders (especially corpus designers) given the task of rating text quality. Then, we evaluate a simple and fully unsupervised method of text quality assessment based on short and very frequent words. Finally, we describe our general approach to the construction of carefully cleansed and non-destructively normalized web corpora. Under this approach, we annotate documents with quality metrics instead of actually removing those documents classified as being of low quality.

1 Introduction

1.1 The Text Criterion

Crawled raw data for web corpus construction contains a lot of documents which are technically in the target language, but which fail as a text. Documents just containing tag clouds, lists of names or products, etc., need to be removed or at least marked as suspicious. Defining the criteria by which the decision to remove a document is made, however, is quite difficult. For instance, many documents contain a mix of good and bad segments and thus represent borderline cases. The decision to systematically remove documents is thus a design decision with major consequences for the composition of the corpus and with potential negative side effects on the distribution of linguistic features. Certain linguistic phenomena might be more or less accidentally underrepresented (w. r. t. the population and/or some specific design criteria) if very long or very short documents are not included, for example. On the other hand, certain lemmas or parts-of-speech might be overrepresented if long word lists or lists of names are not removed, etc. Therefore, while this paper raises mostly technical questions which corpus designers have to care about, we are convinced that linguists working with web corpora should also be aware of how such technical matters have been dealt with.

We first examine how well humans perform given the task of classifying documents as good or bad web corpus documents (Section 2). Then, we introduce and evaluate a completely unsupervised method to classify documents according to a simple but effective metric (Section 3). Finally, we introduce a format for the representation of corpora in which cleanups like boilerplate detection and text quality assessment are not actually executed as deletion. Instead, we keep the potentially bad material and mark it as such (Section 4).

1.2 Context of the Experiment

The work presented here was carried out as part of the construction of the COW2013 corpora, improved versions of the COW2012 corpora (Schäfer and Bildhauer, 2012).[1] The corpora, available in various languages, are all of giga-token (GT) size.[2] Our design goals and the usage scenarios for our web corpora do not allow us to create corpora which are just bags of (very clean) sentences in random order like, for example, the corpora in the Leipzig Corpora Collection (Biemann et al., 2007).[3] We keep whole documents and are generally very careful with all cleanup and normalization steps, simply because the line between noise and corpus material is often difficult to draw. Also, there are many areas of (computational) linguistics

[1] http://www.corporafromtheweb.org/

[2] Currently: Danish 1.5 GT (estimate), Dutch 3.4 GT, English 6 GT (estimate), French 4 GT (estimate), German 9.1 GT, Spanish 1.6 GT, Swedish 2.3 GT.

[3] Cf. also Biemann et al. (2013) for a discussion of different tool chains and their implementation.

for which single sentences are insufficient, such as (web) genre research, information structure, variants of distributional semantics, and even syntax which deals with effects which go beyond single sentences (e. g., the syntax of sentence connectors). Furthermore, one of our future plans is to take uniform random samples from the web by advanced crawling algorithms in order to build small but highly representative web corpora for linguistic web characterization.[4] Although we will always require corpus documents to fulfill minimal linguistically motivated criteria, this general empirically motivated sampling approach does not allow us to filter documents and sentences aggressively, as it would be possible in many more task-oriented settings.

2 Rating Text Quality

2.1 Data Set and Task

Our primary goal in this study was to find out whether corpus designers have clear intuitions about the text quality of web documents, and whether they could operationalize them in a way such that others can reproduce the decisions. Therefore, we randomly selected 1,000 documents from a large breadth-first crawl of the `.uk` TLD executed with `Heritrix` (Mohr et al., 2004).[5] It is the crawl which serves as the basis for our UKCOW2012 and UKCOW2013 corpora. The first 500 documents of the sample were from the initial phase of the crawl, the second 500 from the final phase (after eight days of crawling), when the average quality of the documents is usually much lower (shorter documents, web shops, etc.).[6] The documents were pre-processed with the `texrex` software for HTML stripping, boilerplate removal, code page normalization, etc., and were thus reduced to plain text with paragraph boundaries.[7]

Then, three coders (A, R, S) were given the task of rating each document on a 5-point scale $[-2..2]$ as to how good a corpus document it is.[8] Coders A and R were corpus designers (the second and first author of this paper) with a shared understanding of what kind of corpus they want to build. Coder S was a student assistant who had previously participated in at least three related but not identical rating tasks on the same kind of data, amounting to at least five work days of coding experience.

A series of criteria was agreed upon, the most important being:

- Documents containing predominantly full sentences are good, "predominantly" meaning considerably more than 50% of the text mass (as perceived by the coder).
- Boilerplate material in sentence form is good (*You are not allowed to post comments in this forum.)*, other boilerplate material is bad (*Copyright © 2046 UAC Ltd.*).
- Sentences truncated or otherwise destroyed by some post-processing method are good as long as they are recognizable as (the rest of) a sentence.
- Repetitions of good sentences are good.
- Decisions should not depend on the length of the document, such that a document containing only one good sentence would still be maximally good.
- Non-English material contributes to badness.
- Non-sentence material (lists, tables, tag clouds) contributes to badness.
- However, if a list etc. is embedded in a coherent text which dominates the document, the document is good (prototypically recipes with a substantial amount of instructions).

The scale is interpreted such that 1 and 2 are assigned to documents which should definitely be included in the corpus, -1 and -2 to documents which should not be included, and 0 to borderline cases. In an initial phase, the coders coded and discussed one hundred documents together (which were not included in the final sample) to make results more consistent.[9]

2.2 Results

Table 1 summarizes the results. Despite clear guidelines and the initial training phase, the best

[4] To our knowledge, this has not been done so far. Cf. Chapter 2 of Schäfer and Bildhauer (2013) for an introduction to the problems of uniform sampling from the web and to web characterization. Relevant original papers include Henzinger et al. (2000) and Rusmevichientong et al. (2001).

[5] The data set and the coder data described below can be obtained from the first author.

[6] We will refer to the two subsamples as "early data" and "late data" from now on.

[7] `http://sourceforge.net/projects/texrex/`

[8] There are of course no intrinsically bad or good documents. What we try to measure is the "textiness" of documents, using "goodness" and "badness" as abbreviations for "textiness" and "non-textiness".

[9] It was found in a meta analysis of coder agreement in computational linguistics tasks (Bayerl and Paul, 2011) that training is a crucial factor in improving agreement.

statistic	early 500	late 500	all 1,000
raw	0.566	0.300	**0.433**
κ (raw)	0.397	0.303	0.367
$ICC(C,1)$	0.756	0.679	**0.725**
raw ($r \geq 0$)	0.900	0.762	**0.831**
raw ($r \geq 1$)	0.820	0.674	0.747
κ ($r \geq 0$)	0.673	0.625	**0.660**
κ ($r \geq 1$)	0.585	0.555	0.598
κ ($r \geq 2$)	0.546	0.354	0.498

Table 1: Inter-coder agreement for the text quality rating for 1,000 web documents by three coders; below the line are the results for ratings converted to binary decisions, where $r \geq n$ mean that any rating $r \geq n$ was counted as a positive decision; κ is Fleiss' Kappa and ICC the intraclass correlation.

value ($ICC = 0.756$) on the early 500 documents is mediocre. When the documents get worse in general (and also shorter), the confusion rises ($ICC = 0.679$). Notice also the sharp drop in raw agreement from 0.566 to 0.300 between the early and the late data.

Since Fleiss' κ is not very informative on ordinal data and the ICC is rarely reported in the computational linguistics literature, we also converted the coders' ordinal decisions to binary decisions at thresholds of 0, 1, and 2.[10] The best value is achieved with a threshold of 0, but it is below mediocre: $\kappa = 0.660$ for the whole data set. The value is in fact below the interval suggested in Krippendorff (1980) as acceptable. Even if Krippendorff's interval $(0.67, 0.8)$ is not the final (task-independent) word on acceptable κ values as suggested, for example, in Carletta (1996) an Bayerl and Paul (2011), then 0.660 is still uncomfortably low for the creation of a gold standard. For the binary decisions, the raw agreement also drops sharply from 0.900 to 0.762 between the early and the late material.

It should be noted that coders judge most documents to be quite acceptable. At a threshold ≥ 0 on the 5-point scale, coder A considers 78.4% good, coder R 73.8%, and coder S 84.9%. Still, there is an 11.1% difference between R and S. Positive decisions by R are almost a perfect subset of those by S, however. In total, 73.0% are rated as good by both coders.

We would like to point out that one of the crucial results of this experiment is that corpus designers themselves disagree substantially. Surely, it would be possible to modify and clarify the guidelines, do more training, etc.[11] This would most likely result in higher inter-coder agreement, but it would mean that we operationalize a difficult design decision in one specific way. It has been shown for similar tasks like boilerplate classification that higher inter-coder agreement is possible (Steger and Stemle, 2005). If, however, paragraphs and documents are deleted from the corpus, then users have to agree with the corpus designers on the operationalization of the relevant decisions, or they have to look for different corpora. Our approach is attempt to remedy this situation.

3 Text Badness as the Lack of Function Words

3.1 Summary of the Method

We suggest to use a single criterion in an unsupervised approach to document quality assessment, based on ideas from language identification. In addition to being unsupervised, the approach has the advantage of allowing for very time-efficient implementations. Although the proposed method is arbitrary to a certain degree, it is not a heuristic in the proper sense. As we are going to show, results are quite consistent. Furthermore, considering the degree of arbitrariness involved in human decisions about document quality, we argue against rigorous corpus cleaning and normalization (given the aims and usage scenarios described in Section 1.2) and for non-destructive normalization.

Most approaches to language identification following early papers like Cavnar and Trenkle (1994) and Dunning (1994) use character n-gram statistics. An alternative using short and frequent words is described in Grefenstette (1995). This method (also called the dictionary method) has not been used as prominently as the character n-gram method, but some recent approaches also apply it in the context of normal language identification, e. g., Řehůřek and Kolkus (2009).

[10] Some readers could object that it would have been better to let coders make binary decisions in the first place or redo the experiment in such a way. However, we designed the task specifically because in our earlier informal evaluations and discussions, we had noticed the substantial amount of borderline cases. Using binary decisions or any scale without a middle option would not have captured the degree of undecidability equally well.

[11] Even the word "training" is problematic here, because it is unclear who should train whom.

Clearly, the short word method bears some potential also for text quality detection, because a low frequency of short and frequent words (mostly function words) is typical of non-connected text such as tag clouds, name lists, etc.[12] For the WaCky corpora (Baroni et al., 2009), precompiled lists of words were used, combined with thresholds specifying the required number of types and tokens from these lists in a good document. In Schäfer and Bildhauer (2012), our similar but completely unsupervised method was suggested. It must be mentioned that it only works in an unsupervised manner for web corpora from TLDs with one dominant language. In more complicated scenarios (multilingual TLDs or non-scoped crawls), it has to be combined with normal (i. e., character n-gram based) language identification to pre-filter training documents.[13]

In the training phase, the *n* most frequent word types are calculated based on a sample of documents from the corpus. For each of these types, the weighted mean of its relative frequency in the sampled documents and the corresponding weighted standard deviation are calculated (weighted by the length of the document) as an estimate of the corpus mean and standard deviation. In the production run, these two statistics are used to calculate the normalized deviation of the relative frequency of these *n* types in each corpus document. The more the frequency in the document deviates negatively from the estimated population mean, the worse the document is assumed to be. If the added normalized negative deviation of the *n* types (the "Badness" of the document) reaches a threshold, the document is removed from the corpus. Both in the training and the production run, documents are processed after markup stripping and boilerplate removal.

In practice, we log-transform the relative frequency values because this gave us more consistent results in the initial evaluation. Also, the component value contributed by each of the types is clamped at a configurable value, such that no single type alone can lead to the exclusion of a document from the corpus. This was motivated by the fact that, for example, in many languages the personal pronoun for *I* is among the top ten types, but there are certain kinds of documents in which it does not occur at all because self-reference is sometimes considered inappropriate or unnecessary. The clamping value was set to 5 for all experiments described here. A short-document bias setting is also available, which reduces the Badness of short documents (because relative frequencies show a higher variance in short documents), but we currently do not use it in evaluations and in production runs.

3.2 Evaluation of Type Profiles

We use the ten most frequent types to generate frequency profiles, since the ten most frequent types usually make up for more than one fifth of the tokens in documents/corpora (Baroni, 2008), and they can be considered to have a reasonably domain-independent distribution. Figure 1 shows how the log-transformed weighted arithmetic mean and the corresponding standard deviation for the 10 most frequent types develop while training the DECOW2012 reference profile trained on 1,000 documents from the beginning of the crawl ("early profile"). As expected, both the mean and the standard deviation are relatively stable after 1,000 documents. The occasional jumps in the standard deviation (most remarkably for *und*) are caused by very long documents (sometimes over 1 MB of text) which thus receive very high weights. Future versions of the software will include a document size pre-filter and the option of using different profiles for documents of different length to smooth this out. However, given the evaluation results in Section 3.4, we think these additional mechanisms are not crucial.

3.3 Distribution of Badness Values

Next, we look at the distribution of the Badness values under realistic corpus processing scenarios. We used the early DECOW2012 profile described in Section 3.2 to calculate Badness values for a large number of early documents (2.2 GB HTML data; 27,468 documents), i. e., documents from the same phase of the crawl as the ones used for training the profile.[14] We did the same with early UKCOW2012 data (2.2 GB HTML data; 32,359 doc-

[12] In this sense, the method is, of course, not arbitrary, but based on quite reasonable theoretical assumptions about the distributions of words in texts.

[13] We have successfully used available n-gram-based language-identification in a task-specific crawling scenario (Barbaresi, 2013) and are planning to integrate all methods into one piece of software eventually.

[14] We use "early/late data" for "data from the early/late phase of the crawl" and "early/late profile" for "profile trained on a sample of the documents from the early/late phase of the crawl" from now on.

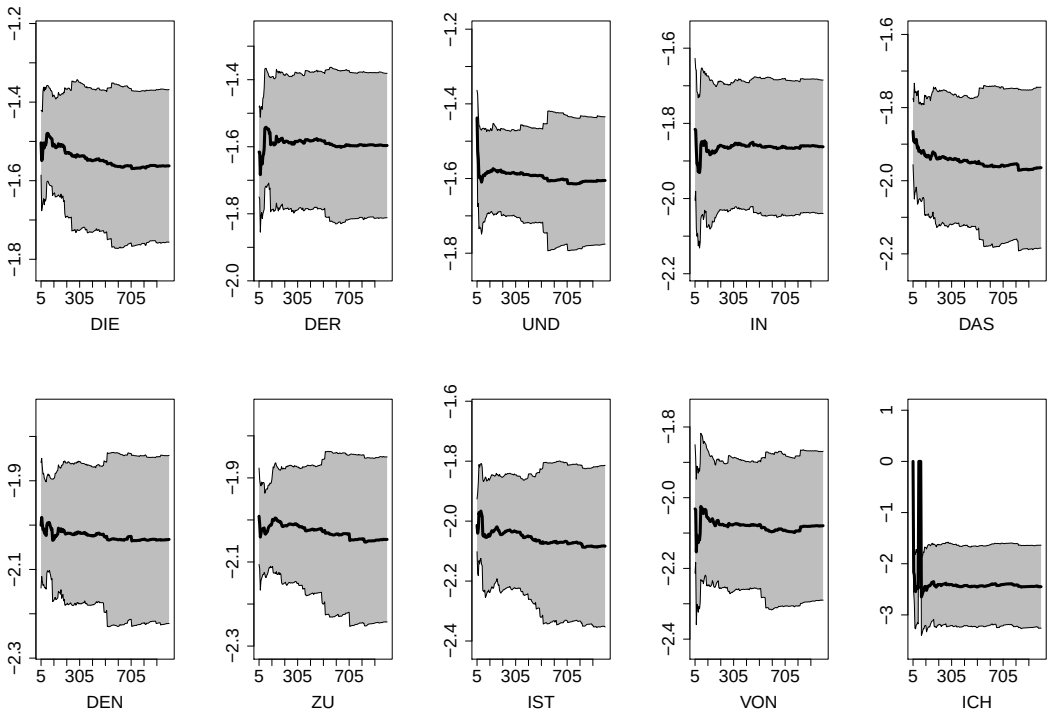


Figure 1: Development of the reference profile for DECOW2012 on early crawl data ("early profile") while training; x-axis: number of documents used for training; y-axis: log10-transformed weighted arithmetic mean of the respective type's frequency in the training documents; gray areas mark 1 standard deviation above and below the mean; the 10 most frequent types after 1,000 documents.

uments) and an early profile.[15] Figure 2 shows the resulting distribution of Badness values for documents above certain byte lengths.

In the early phase, the UKCOW2012 crawl found more short documents compared to the early phase of the DECOW2012 crawl, namely 4,891 (17,81%) documents more for 2.2 GB of raw data. The mean document length is therefore lower for UKCOW2012. This generally lower document length probably explains the different shape of the distribution, i. e., the higher overall Badness of the UKCOW2012 documents. In both cases, however, there are a lot of very bad documents (Badness=50) at short byte lengths. They are typically those documents which are completely or at least almost empty after boilerplate removal. For the following evaluations, we therefore removed all documents up to a length of 200 B.

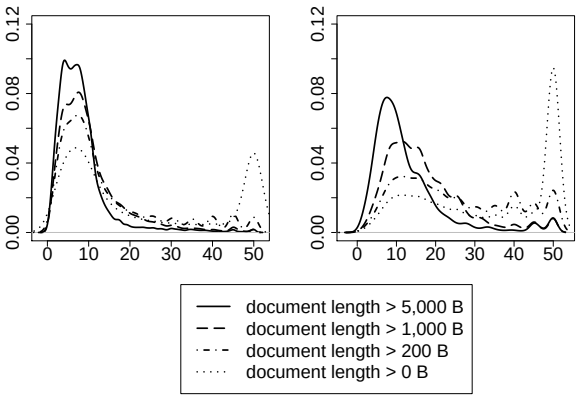


Figure 2: Density estimates for the distribution of Badness scores for the early profile on early data depending on document length; left: DECOW2012 ($n = 27,468$), right: UKCOW2012 ($n = 32,359$); x-axis: Badness score/threshold; y-axis: distribution density.

3.4 Comparison of Profiles

We now look at the question of whether profiles created from different samples have radically different effects. To this end, comparisons are made between the effects of profiles created *with* early and late data *on* early and late data, respectively.

[15] The UKCOW2012 early data here is a superset of the documents used in the coding task described in Section 2.

Figure 3 plots (for documents longer than 200 B) the proportion left over by early profiles on early data, etc.

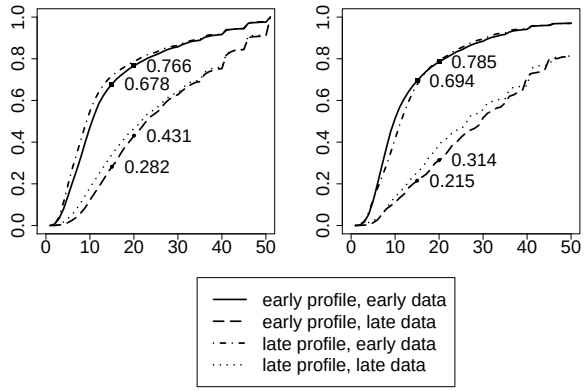


Figure 3: Effect of different profiles in terms of the proportion of documents left over at certain Badness thresholds (≡ cumulative density distribution of Badness values) for all documents longer than 200 B; left: DECOW2012 ($n = 27,468$ early; $n = 60,565$ late); right: UKCOW2012 ($n = 32,359$ early; $n = 34,879$ late); x-axis: Badness score/ threshold; y-axis: proportion of documents left in the corpus; values at Badness 15 and 20 for the early profile are given in the graphs.

In the case of DECOW2012, the early data sample contains documents which are on average 2.2 times longer than those in the late data sample. Profiles trained on documents from two such different samples would be likely candidates for having different effects. Surprisingly, the different profiles have rather negligible effects. For the early DECOW2012 data, the early profile leaves 76.6% of the document in the corpus, while the late profile leaves 78.6%, a difference of no more than 2%. On late data, it is 43.1% (early profile) and 46.4% (late profile). For the UKCOW2012 early data, it is 78.5% (early profile) vs. 79.2% (late profile) and for the late data 31.4% (early profile) and 39.1% (late profile). As expected, due to higher variance in the training data (which is mostly due to shorter document length), late profiles are more permissive, but the differences are not drastic. Figure 4 plots the raw agreement of the profiles on the early and the late data set at Badness thresholds from 1 to 50. It shows that the major difference is the reduced strictness of the late profiles on the early data, but mainly below thresholds of roughly 15 or lower.

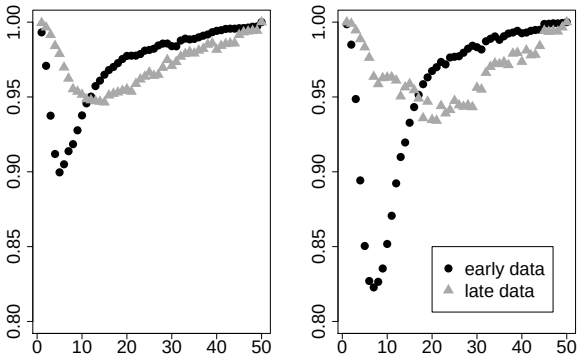


Figure 4: Profile comparison in terms of raw agreement between the profiles at thresholds [1..50]; left: DECOW2012; right: UKCOW2012; x-axis: thresholds; y-axis: proportion of identical decisions of the early and the late profile at the given threshold.

Finally, Figure 5 confirms the general picture. For the two TLDs (`.de` and `.uk`), it plots the Badness values calculated by the early and the late profiles on the two data sets. Each of the four plots corresponds to one data set (early or late) from one of the TLD crawls, and it compares the two profiles w. r. t. those data sets. Each dot represents a document, and it is positioned to show the Badness value assigned to that document by the late profile (x) and the early profile (y).

The linear models on the data show quite a strong correlation between the Badness scores assigned by the two profiles. The intercepts are higher for late data compared to earlier data (DECOW2012: early data 1.028, late data 2.194, UKCOW2012: early data 0.994, late data 3.741), showing again that the early profiles are more sensitive/strict than the late profiles.

Why the UKCOW2012 data is worse in general is impossible to ascertain. Since the seed URLs were collected in a similar way for both crawls, and the crawler software was configured in exactly identical ways, the difference is most likely a symptom of the unpredictable biases brought about by unselective Breadth-First Search.

3.5 Avoiding Impossible Decisions

So far, we have shown that deciding whether a document contains mostly text (as opposed to non-

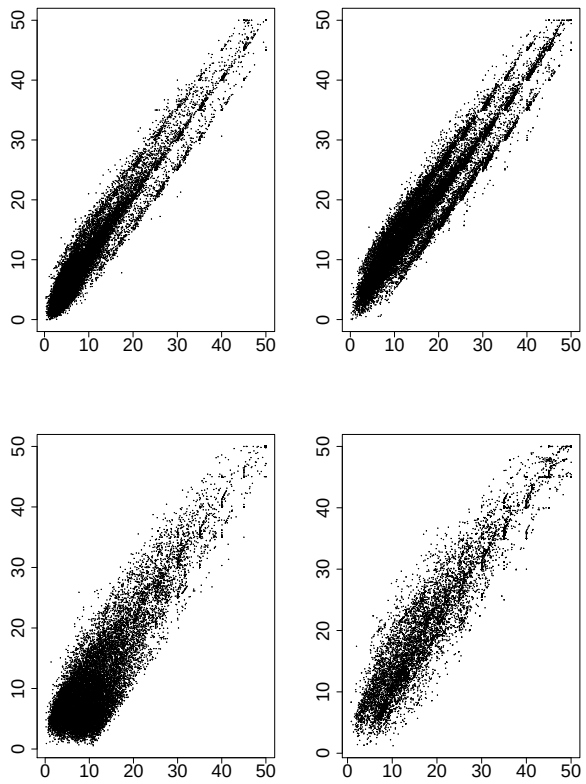


Figure 5: Comparison of profiles; top: DECOW2012; bottom: UKCOW2012; left: early data; right: late data; x-axis: late profile; y-axis: early profile; LM top left (DECOW2012 early data): intercept=1.028, coefficient=0.980, R^2=0.988; LM top right (DECOW2012 late data): intercept=2.194, coefficient=0.952, R^2=0.982; LM bottom left (UKCOW early data): intercept=0.128, coefficient=0.994, R^2=0.9701; LM bottom right (UKCOW late data): intercept=3.741, coefficient=0.930, R^2=0.970; artefacts around Badness increments of 5 result from the clamping (to 5) of the values which are added up to calculate Badness.

text material) is a task which leads to substantial disagreement between humans. Furthermore, we have argued that the lack of short and otherwise highly frequent words can be measured easily and with consistent results for the kind of data in which we are interested. However, if we want to use the Badness score as a document filter, then there still remains a threshold to be determined, i. e., a score above which documents are excluded from the corpus. We now discuss how such a value should

	prec	rec	F1	correct	baseline
S	0.914	0.959	0.936	0.888	0.849
A	0.856	0.973	0.911	0.851	0.781
R	0.808	0.976	0.884	0.811	0.738

Table 2: Performance of the Badness algorithm as a classifier evaluated against the human coder decisions; thresholds chosen to produce the maximal possible agreement with any coder (which is coder S): coder threshold 0; Badness threshold 35; raw agreement of the human coders is 0.831 at these settings (Fleiss' $\kappa = 0.660$).

be chosen by comparing the Badness scores for the 1,000 UKCOW2012 documents from the experiment described in Section 2 with the coders' decisions.

We searched for the best match between Badness scores and coder decisions and found that if we keep documents rated by coder S (the least strict coder) as 0 or better, then setting the Badness threshold to 35 results in a proportion of correct predictions of 0.888, cf. Table 2. This is the best achievable value for any coder and any Badness threshold with the data from our coding task.

For a (hypothetical) gold standard based on coder S, the Badness score method achieves a precision, a recall, and an F1 score of well over 0.9. Of course, since the baseline ("keep all documents") is quite high, this means an increase in accuracy of only 0.039 (roughly 4%) compared to the baseline. At the same time, Table 2 shows that at the optimal settings for coder S, the methods achieves a precision below 0.9 (more bad documents remaining in the corpus) relative to the decisions by the other coders. Still, even for the strictest coder (R), precision is above 0.8. The recall, however, is generally excellent.

We suggest that the best lesson to learn from these results is that corpus designers should not make too many destructive design decisions, ideally none at all. If we keep all documents accepted as good enough for corpus construction by the most tolerant coders, then all users can be sure that the material in which they are interested is still contained in the corpus (near-perfect recall for everyone). If, in addition to this, we annotate the documents with (ideally several) metrics like the Badness score, corpus users can decide to use more or less clean and/or good documents when making queries or generating statistics from

the corpus. In other words, corpus users should be put in a position to decide how important precision and recall are for their purposes. This is currently our general strategy, and we summarize it in more detail in the next and final section.

4 Achievements, Further Research, and Corpus Formats

As it was said in Section 1.2, our ultimate target in web corpus construction is the creation of highly representative samples from the population of web documents with the least possible error introduced through post processing and normalization. Therefore, in addition to working on improved crawling methods, we let users decide how strictly they want to filter potential noise. The measures which we have already implemented and used for the construction of the UKCOW2012 corpus (which, however, was still crawled using a Breadth-First Search) are the annotation of documents with Badness scores and the annotation of paragraphs with values indicating the likelihood that they are boilerplate. Since we are currently in the stage of evaluation of diverse methods, we still removed documents with a Badness of 15 or higher (not 35, as suggested in Section 3.5), and we removed paragraphs with a certain likelihood of being boilerplate (although not as strictly as in earlier COW2012 corpora). For COW2013, we are planning to keep all paragraphs and all documents below a Badness of 35.

Since we use the IMS Open Corpus Workbench (CWB) for corpus access, we needed to encode the Badness and boilerplate scores in a way such that they can be used in CQP queries.[16] Adding the raw numeric values to structural attributes is not a feasible way of doing this, because CWB would basically treat them as factors, not enabling queries restricted by arithmetic conditions on those values. In other words, querying for documents with a Badness smaller than r_1 and greater than r_2, etc., is impossible. We therefore encode the values as single alphabetic characters between *a* (best) to maximally *z* (worst). Badness values were encoded in increments of 2, such that $[0,2)$ is encoded as *a*, $[2,4)$ as *b*, etc. For example, restricting the search to documents with a Badness of 10 or better can be achieved by specifying the regular expression `[a-e]` for the Badness annotation layer.

[16] `http://cwb.sourceforge.net/`

Of course, the amount of data increases considerably with this highly non-destructive approach to post processing and normalization. From an empirical point of view, this simply is not a valid counter-argument. What is more, it is quite feasible to construct giga-token corpora in such a way on modern hardware without serious performance penalty, as we have demonstrated with UKCOW2012. Furthermore, given that uniform random sampling allows for smaller samples in order to achieve representativeness, the effect of non-destructive cleansing and normalization on corpus size can be compensated for in the long run by using smaller samples in the first place. While very huge (and traditionally cleaned/normalized) corpora in the region of several 10^7 tokens (Pomikálek et al., 2012) are surely very useful, for some applications in empirical linguistics, better is better, and bigger is not necessarily better.

Acknowledgments

We would like to thank Sarah Dietzfelbinger for her participation in the coding task. Also, we would like to thank three anonymous WaC 8 reviewers for their helpful comments. Felix Bildhauer's work was funded by the *Deutsche Forschungsgemeinschaft* through the *Sonderforschungsbereich* 632, project A6.

References

Adrien Barbaresi. 2013. Crawling microblogging services to gather language-classified URLs. Workflow and case study. In *Proceedings of ACL Student Research Workshop*, Sofia. To appear.

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

Marco Baroni. 2008. Distributions in text. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, pages 803–822. Walter de Gruyter, Berlin.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

C. Biemann, G. Heyer, U. Quasthoff, and M. Richter. 2007. The Leipzig Corpora Collection - Monolingual corpora of standard size. In *Proceedings of Corpus Linguistics 2007*, Birmingham, UK.

Chris Biemann, Felix Bildhauer, Stefan Evert, Dirk Goldhahn, Uwe Quasthoff, Roland Schäfer, Johannes Simon, Leonard Swiezinski, and Torsten Zesch. 2013. Scalable construction of high-quality web corpora. *Special issue of JLCL*. In prep. The list of authors is preliminary and might reflect neither the order nor the actual list in the printed version.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

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

Ted Dunning. 1994. Statistical identification of language. Technical Report MCCS-94-273, Computing Research Laboratory, New Mexico State University.

Gregory Grefenstette. 1995. Comparing two language identification schemes. In *Proceedings of the 3rd Internation conference on Statistical Analysis of Textual Data (JADT 1995)*, pages 263–268, Rome.

Monika R. Henzinger, Allan Heydon, Michael Mitzenmacher, and Marc Najork. 2000. On near-uniform URL sampling. In *Proceedings of the 9th International World Wide Web conference on Computer Networks: The International Journal of Computer and Telecommunications Networking*, pages 295–308. North-Holland Publishing Co.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills.

Gordon Mohr, Michael Stack, Igor Ranitovic, Dan Avery, and Michele Kimpton. 2004. Introduction to Heritrix, an archival quality web crawler. In *Proceedings of the 4th International Web Archiving Workshop (IWAW'04)*.

Jan Pomikálek, Miloš Jakubíček, and Pavel Rychlý. 2012. Building a 70 billion word corpus of English from ClueWeb. In *Proceedings of LREC 08*, pages 502–506.

Paat Rusmevichientong, David M. Pennock, Steve Lawrence, and C. Lee Giles. 2001. Methods for sampling pages uniformly from the World Wide Web. In *In AAAI Fall Symposium on Using Uncertainty Within Computation*, pages 121–128.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul. ELRA.

Roland Schäfer and Felix Bildhauer. 2013. *Web Corpus Construction*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool, San Francisco.

Johannes Steger and Egon Stemle. 2005. Krdwrd architecture for unified processing of web content. In Iñaki Alegria, Igor Leturia, and Serge Sharoff, editors, *Proceedings of the Fifth Web as Corpus Workshop (WAC5)*, pages 63–70, San Sebastián. Elhuyar Fundazioa.

Radim Řehůřek and Milan Kolkus. 2009. Language identification on the web: Extending the dictionary method. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 5449 of *Lecture Notes in Computer Science*, pages 357–368. Springer Berlin Heidelberg.

Developing a User-based Method of Web Register Classification

Jesse Egbert
Northern Arizona University
Box #6032
Flagstaff, AZ 86011
Jesse.Egbert@nau.edu

Douglas Biber
Northern Arizona University
Box #6032
Flagstaff, AZ 86011
Douglas.Biber@nau.edu

Abstract

This paper introduces a new grant-funded initiative to develop a comprehensive linguistic taxonomy of English web registers. We begin with an overview of the goals, methods, and current status of the project. However, we focus mostly on a detailed discussion of the methods used to develop a user-based register classification rubric, and a presentation of the results obtained to date coding a large corpus of web documents for their register categories.

1 Introduction

The World Wide Web is a tremendous resource of information which is growing at an accelerated rate. The identification of register (or genre) is particularly important for natural language processing (NLP) applications in computational linguistics, improving the performance of word disambiguation software, taggers, parsers, and information retrieval tools. Linguists have also recently begun to use the web as a corpus for studies of linguistic variation and use. However, the unique nature of the different types of language used on the web remains unclear. Without a clear understanding of the linguistic variability of internet texts we are severely limited in our ability to use this powerful resource for linguistic and NLP research.

In order to better understand the language of the internet, it needs to be systematically classified into registers/genres. In recent years there has been a surge of interest in Automatic Genre Identification (AGI), which is a computational method of using a wide range of descriptors to automatically classify web texts into genre classes. Several AGI approaches have achieved high accuracy rates (e.g. Sharoff, et al., 2010). However, these AGI corpora are often manually classified by only one person (see Sharoff et al., 2010). In addition, the nature of the corpora used for tests of AGI raises questions about the potential accuracy of these models with real internet texts. For example, AGI researchers seldom know whether the sample in a given corpus represents the full population of internet texts or whether the texts within a given genre class represent the variability of the descriptors included in the model (see Santini & Sharoff, 2009).

An even more fundamental question regarding the usefulness of web genre corpora concerns the reliability of manual genre classification. As mentioned above, most corpora used to test AGI models are sub-divided into genre classes, but this is often done by only one person (Sharoff et al., 2010). Calculating reliability among multiple raters may seem unnecessary if an "expert in genre-related research" plans to code all texts (Rehm et al., 2008). However, the few cases where inter-rater reliability is reported have shown that it tends to be quite low, even for linguists. This is especially true for corpora comprised of randomly extracted web texts (Sharoff et al., 2010). Given the problems that 'experts' have identifying web genre categories, it is not surprising that non-expert web users also vary in their understanding of genre labels (see Crowston et al., 2010), and that reliability among users is often unacceptably low (Rosso & Haas, 2010).

2 Overall project goals and methods

In order to address the aforementioned research gaps, we set out to answer the following questions:

1. What are the web register distinctions recognized by non-expert internet users?
2. To what extent can non-expert raters reliably classify web texts into those register categories?
3. What is the distribution of English-language registers on the web?

We use the term 'register' rather than 'genre' to refer to the textual distinctions focused on in our investigation, following the framework developed in Biber and Conrad (2009).

The first step in answering our research questions was to create a large corpus of internet language (c. 1 million web pages) by using the results of Google searches of highly frequent English 3-grams (e.g., *is not the, and from the*). The use of n-grams as search engine seeds is an approach that has been used in the past by many web-as-corpus scholars (see, e.g., Baroni & Bernardini, 2004; Baroni et al., 2009; Sharoff, 2005; 2006). The search results were then downloaded using HTTrack (http://www.httrack.com), and JusText (http://code.google.com/p/justext) was used for HTML scrubbing and boilerplate removal.

The majority of the present talk will focus on the methods used to develop a comprehensive web register classification rubric and apply that framework to a large corpus of web documents. In Section 3 we will summarize the various steps we took in order to develop and pilot a new web register classification instrument. Section 4 presents the reliability results of a large-scale study which was carried out using this instrument. Section 5 presents frequency information for the general registers, sub-registers, and hybrid registers identified in the study described in Section 4.

3 Developing and piloting a register classification instrument

In order to answer the first research question, "What are the web register distinctions recognized by non-expert internet users?" we set out to develop an instrument that can be used by non-expert web users to classify web documents into web register categories.

After reviewing a large number of studies where register/genre palettes were developed, we chose to follow Rehm et al's (2008) suggestion to begin with the 78 categories that resulted from a wiki-based collaboration among web-as-corpus experts (http://www.webgenrewiki.org/). Based on our own previous experience with register analysis, we grouped those 78 categories into 8 general registers (e.g., description, opinion, non-fiction narrative), with several sub-registers within each top-level category (e.g., opinion: opinion blogs, editorials, reviews, advice).

We then embarked on a series of pilot studies to refine our framework, with the overall goal of including all web register distinctions that are recognized by end-users and able to be applied with high reliability in practice. Basic descriptive information and results from each of the ten pilots is displayed in Table 1 below.

Table 1. Overview of the ten pilot studies

Round	URLs	Raters	Results		Post-analysis modifications
			GR	SR	
Stage I. Rubric with descriptions					
1	25	2	72% sa	---	-transformed rubric into flowchart -Added examples of sub-register categories and presented them in order of frequency
Stage II. Flowchart with examples					
2	25	2	64% sa	---	-Added brief explanations of each decision in the flowchart -Texts that are more than 50% quotes are considered spoken -Texts that are more than 50% reader comments are considered discussion -Created a distinction between technical and non-technical discussions
3	25	2	68% sa	---	-Reader comments are to be noted rather than considered discussion -Heavy quotes are to be noted rather than considered spoken
Stage III. Online survey					
4	25	2	68% sa	---	-Added 3-4 examples after each option -Added Informational persuasion as a general register -Modified the wording of options to improve clarity -Created a page for raters to select a sub-register category for each text -Increased number of raters to 3 in order to increase accuracy of agreement results
5	25	3	3/3: 59% ≥2/3: 100%	3/3: 41% ≥2/3: 82%	-Increased the number of example sub-registers for the options on each page -Added a drop-down menu on the first page with all sub-register options to be used only when the rater is certain about the register of the text

					-Reordered the options on several pages so the most used options are at the top
6	25	3	3/3: 55% ≥2/3: 100%	3/3: 41% ≥2/3: 82%	-Added a more comprehensive list of sub-registers to the options on each page -Began calculating agreement for 2/3 and 3/3 rater agreement -Added 'or co-authors' to the one author option in order to allow for non-discussion texts written by more than one person -Removed the drop-down menu option because it seemed to cause lower agreement
7	25	3	3/3: 56% ≥2/3: 92%	3/3: 49% ≥2/3: 84%	-Increased number of raters to 5 in order to gather data based on a majority rather than perfect agreement -Added a drop-down menu for the two most common sub-registers
8	50	5	5/5: 38% ≥4/5: 62% ≥3/5: 94%	5/5: 26% ≥4/5: 48% ≥3/5: 72%	-Reduced number of raters to 4 in order to identify hybrids by allowing for ties -Added an option for 'not enough text to rate' -Eliminated the distinction between technical and non-technical descriptions because of low agreement
9	100	4	4/4: 33% ≥3/4: 68% 2-2 tie 12%	4/4: 18% ≥3/4: 53% 2-2 tie 16%	-Added an option for 'website not found' -Increased the number of sub-registers in the drop-down menu to 4 -Eliminated two pages by merging multiple options into single survey pages
10	1k	4	4/4: 34% 3/4: 29% 2-2 tie 11% 2-1-1 19_%	4/4: 18% 3/4: 25% 2-2 tie 8% 2-1-1 10%	-Eliminated one page by merging multiple options into single survey pages (personal narrative, fictional narrative, factual narrative → past narrative)

Note: GR: general register; SR: sub-register; sa: simple agreement

Our first step toward a reliable register classification instrument was to create a rubric that contained descriptions of each of the general register categories and examples of text types that would be classified in each (see Table 1, Stage I). After the first round of piloting, we determined that the classification rubric was too time-consuming and cumbersome to use, so we adapted it into a visual flowchart that guided the rater through a series of simple choices until they had arrived at most appropriate register category for the web document (see Table 1, Stage II).

In order to further improve the speed and ease of the rating task, we transformed the flowchart into a computer-adaptive online survey (see Table 1, Stage III). This survey was designed to present one set of multiple choice options to the raters on each screen. For example:

The main purpose of this text is to…

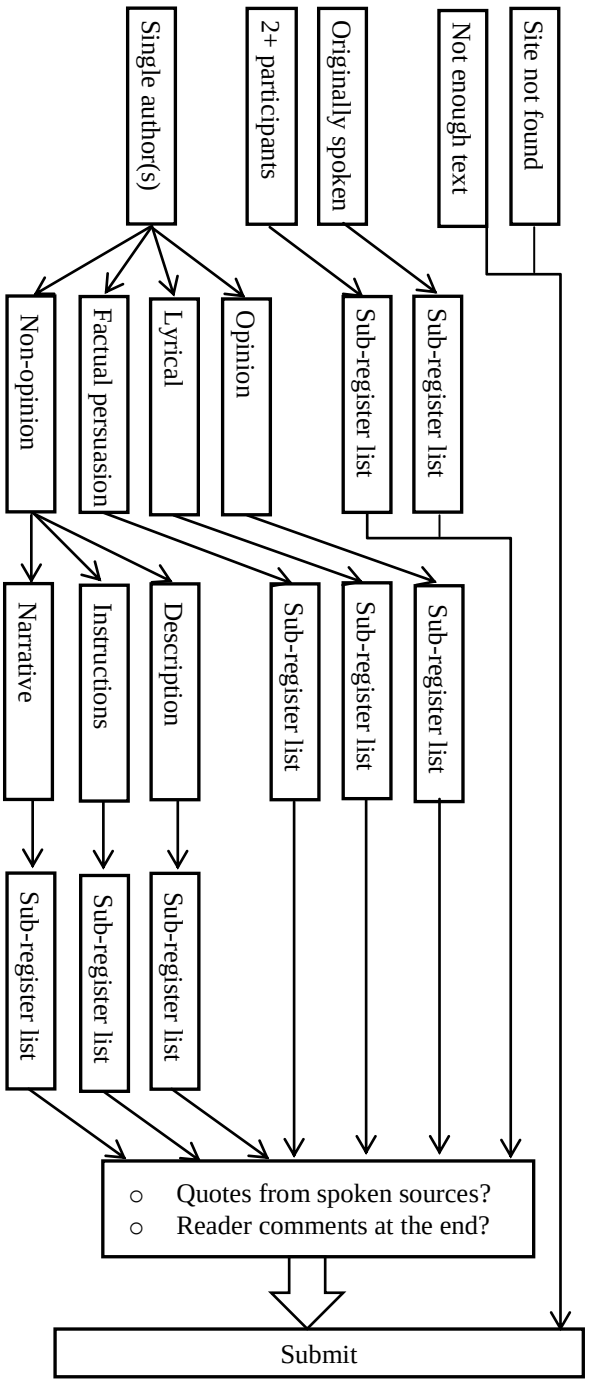
- narrate or report on PAST EVENTS
- describe or explain INFORMATION
- explain HOW-TO or INSTRUCTIONS

Based on their previous choices, the raters were guided through a series of 2-6 pages until they had reported enough information for us to assign a general register and sub-register label to each web document. The flowchart in Figure 1 is a visual representation of the final version of the computer-adaptive survey which was used for the present study.

An in-depth analysis of results from the online survey revealed a wide range of variability in the agreement for the individual general registers and sub-registers. This led to further refinements to the categories and the structure of the survey in order to improve agreement. Additionally, these analyses revealed a number of hybrid categories on the general and sub-register levels. In other words, user classification for webpages that achieved low reliability were often split between two categories, and some of these 'hybrid' categories emerged repeatedly across multiple texts.

The discovery of hybrid web texts was something we anticipated based on the findings of previous research (see, e.g., Santini, 2007; 2008; Vidulin, Luštrek, & Gams, 2009). In order to address register hybrids, we decreased the number of raters from five to four for the next round of data collection. This allowed for the possibility of a 2-2 tie between two register categories.

Figure 1. Flowchart displaying possible paths through the computer-adaptive survey



4 Reliability of web register classification

The second research question for our study was: "To what extent can non-expert raters reliably classify web texts into register categories?" To answer this question, we carried out a pilot study with 4 internet raters coding 1,000 webpages.

This tenth pilot study, along with pilot studies 8 and 9, was administered to workers through Mechanical Turk (MTurk), an Amazon-based online crowdsourcing company. It should be noted that the earlier pilot rounds contained webpages from many geographic locations, but the webpages used from this point forward were drawn only from texts produced in the following five countries: United States, Great Britain, Canada, Australia, and New Zealand. This was done in order to reduce variability in the linguistic and situational characteristics of the web texts classified by participants. We also added two important options to the survey in this stage: (a) a 'website not found' option, and (b) a 'not enough text to classify' option.

A 7-minute interactive video tutorial was created in order to introduce the purpose of the project and provide training to the MTurk workers. After participating in the tutorial, a practice webpage was provided to each potential rater. Those who correctly classified the practice webpage were subsequently authorized to rate additional webpages.

Based on the results of the ninth pilot study reported in Section 3, for round 10 we decided to quantify frequent 3-way hybrids, in addition to 2-2 ties. These 2-1-1 patterns (e.g., Narrative+Narrative+Opinion+Description), along with the 2-2 hybrid patterns, were counted only if the particular pattern occurred more than five times. Approximately 3.6% of the pages were not found and about 3.3% of the pages were labeled as not having enough text to rate. The frequency data in Table 2 includes only the webpages that were actually coded for register (n = 931).

Table 2. Results for 931 webpages using MTurk*

General Registers				
4 agree	3 agree	2-2 hybrid	2-1-1 hybrid	No agree
315	269	104	173	70
33.8%	28.9%	11.1%	18.6%	7.6%
Sub-registers				
4 agree	3 agree	2-2 hybrid	2-1-1 hybrid	No agree
171	231	73	90	366
18.3%	24.8%	7.8%	9.8%	39.3%

*69 texts were not rated (36-'websited not found; 33-'not enough text')

It can be seen that a majority agreed on the register category for 62.7% of the texts. About 30% of all webpages were classified as hybrids. These data show that our method made it possible to reliably classify over 92% of the random sample of webpages with text into a meaningful general register or general register hybrid category.

On the sub-register level, a majority of the raters was able to agree on about 43% of the web texts. A much smaller proportion of the texts fell into a hybrid sub-register category (17.5%). Together, these data show that about 61% of the webpages could be successfully categorized into a specific sub-register category or a sub-register hybrid.

On the whole, the results presented here show that non-expert web users can, to a large degree, use a computer-adaptive web register classification survey to reliably classify webpages into general registers, sub-registers, and register hybrids.

5 The distribution of English language registers on the web

The third research question for our study was "What is the distribution of English language registers on the web?" In order to answer this question, we present frequency information for the general registers and sub-registers included in pilot study 10 (see Tables 3 and 4). This frequency information is based on the web documents for which a majority of the raters agreed (n = 402). The general registers in Table 3 and the sub-registers in Table 4 are presented in order of frequency, and both tables contain frequency and percentage information.

Table 3. Frequency information for general register categories

General Register	#	%
Narrative	**135**	**33.6**
Opinion	**95**	**23.6**
Description	**67**	**16.7**
Discussion	**54**	**13.4**
Lyrical	**18**	**4.5**
How-to/Instructional	**16**	**4.0**
Informational Persuasion	**10**	**2.5**
Spoken	**7**	**1.7**

Table 4. Frequency information for sub-register categories

Register	#	%
Narrative	**135**	
News report/blog	*99*	*73.3*
Sports report	*19*	*14.1*
Personal/diary blog	*7*	*5.2*
Historical article	*4*	*3.0*
Short story	*3*	*2.2*
Novel	*2*	*1.5*
Biographical story/history	*1*	*0.07*
Joke	*0*	*0*
Magazine article	*0*	*0*
Memoir	*0*	*0*
Obituary	*0*	*0*
Other factual narrative	*0*	*0*
Other fictional narrative	*0*	*0*
Other personal narrative	*0*	*0*
Travel blog	*0*	*0*
Opinion	**95**	
Opinion blog	*57*	*60.0*
Review	*23*	*24.2*
Advice	*9*	*9.5*
Religious blog/sermon	*5*	*5.3*
Self-help	*1*	*1.1*
Advertisement	*0*	*0*
Letter to the editor	*0*	*0*
Description	**67**	
Description of a thing	*34*	*50.7*
Description of a person	*9*	*13.4*
Research article	*7*	*10.4*
Abstract	*5*	*7.5*
Legal terms and conditions	*4*	*6.0*
FAQ about information	*2*	*3.0*
Encyclopedia article	*2*	*3.0*
Informational blog	*2*	*3.0*
Course materials	*1*	*1.5*
Technical report	*1*	*1.5*
Other	*0*	*0*
Discussion	**54**	
Question/answer forum	*46*	*85.2*
Other forum	*7*	*13.0*
Other discussion	*1*	*1.8*
Reader/viewer responses	*0*	*0*
Lyrical	**18**	
Song lyrics	*17*	*94.4*
Other	*1*	*5.6*
Poem	*0*	*0*
Prayer	*0*	*0*
How-to/Instructional	**16**	
How-to	*13*	*81.3*
Technical support	*2*	*12.5*
Recipe	*1*	*6.2*

Instructions	*0*	*0*
FAQ about how to do something	*0*	*0*
Other	*0*	*0*
Informational Persuasion	**10**	
Description with intent to sell	*8*	*80.0*
Persuasive article or essay	*2*	*20.0*
Editorial	*0*	*0*
Other	*0*	*0*
Spoken	**7**	
Interview	*5*	*71.4*
Formal speech	*1*	*14.3*
Transcript of video/audio	*1*	*14.3*
Other	*0*	*0*
TV/movie script	*0*	*0*

The most common general internet register is Narrative. The frequency results for the sub-registers showed that nearly 90% of the texts in this general register were classified as either News report/blogs or Sports reports. The next most frequent general register is Opinion, of which 60% were classified as Opinion blogs and 24% were classified as Reviews. The general register of Description comprised almost 17% of the sample, with Description of a thing and Description of a person making up nearly two-thirds of the sample. The Discussion general register was also used relatively frequently, and the vast majority of these texts were classified as Question/answer forums. The Lyrical, How-to/Instructional, Informational Persuasion, and Spoken general registers each occurred much less frequently than the first four. However, it is clear that these general registers each comprise one or two important sub-register categories.

While some of these general registers and sub-registers are very similar to traditional print registers (e.g., News reports, Sports reports, Reviews, Research articles, Song lyrics), many of them are unique to the domain of the internet. For example, the sub-registers of Personal/diary blogs and Opinion blogs, as well as the general register of Discussion, can only be found on the internet. Furthermore, many of the registers that appear to be traditional are actually quite different from their printed, non-internet counterparts. This is due to several factors, including the relative ease of 'publishing' on the internet and decreased attention to pre-planning and editing common in many internet registers.

As mentioned above, in addition to the webpages for which a register category was agreed upon by a majority of the participants, many of the webpages were classified into hybrid registers based on the varied registers assigned to them by the participants. Some patterns emerged from an analysis of the frequency data for these hybrid combinations. Table 5 displays the counts for the top five 2+2 general register hybrids. It is apparent that Opinion and Description are prolific members of these hybrid combinations.

Table 5. Five most frequent general register 2+2 hybrid combinations

Hybrid Combination (2+2)	Count
Description + Narrative	43
Narrative + Opinion	27
Description + Opinion	17
Informational Persuasion + Opinion	11
Description + Informational Persuasion	6

In addition to webpages that resulted in a 2-2 tie between two registers, we also counted cases where a webpage was coded with three different register categories, resulting in a 2-1-1 split. Although it is possible that some of these occurred by chance alone, the results presented in Table 6 suggest that these patterns represent actual underlying register hybrids.

Table 6. Five most frequent general register 2+1+1 hybrid combinations

Hybrid Combination (2+1+1)	Count
Narrative + Opinion + Description	56
Description + Informational Persuasion + Opinion	40
Description + Informational Persuasion + Narrative	28
Informational Persuasion + Narrative + Opinion	24
Description + How-to/Instructional + Opinion	15

For example, the most frequent 3-way hybrid is Narrative + Opinion + Description. An example of one such webpage was labeled a News report/blog (2 raters), a Description of a person (1 rater), and an Opinion blog (1 person). The title of this text, which is "On the road: Bradley Wiggins and Team Sky have made Tour de France history – it's been emotional," is suffi-

cient to demonstrate the triad of characteristics recognized by the four raters. This text is a blog post that recounts a recent news story (Narrative) that describes a team of athletes (Description) from the perspective of the author (Opinion).

An additional factor that has not yet been addressed is the common option for readers to comment below an internet text, such as a News report/blog or a Review. In an effort to differentiate between standalone texts and those with reader comments, participants were asked to check a box at the end of the survey if the text contained reader comments. The results show that 234 (25.1%) of the 931 webpages that received ratings contained at least some reader comments at the end of the text. Table 7 displays the distribution and proportion of texts with reader comments in each of the general register categories.

Table 7. Frequency information for texts containing reader comments

Register	Count	Percent
Narrative	87	37.2
Opinion	86	36.8
Description	37	15.8
Informational Persuasion	12	5.1
How-to/Instructional	8	3.4
Lyrical	4	1.7
Spoken	0	0
Discussion	0	0
Total	**234**	**100**

6 Conclusion

The user-based register classification methodology outlined here is novel in a number of ways. This is the first study we know of that takes a bottom-up approach to developing a register framework (or 'genre palette') using a random sample of internet texts. Additionally, the number of text samples coded is unprecedented in previous research. Finally, this is one of the first efforts to develop a web register classification rubric designed for use by non-expert web users.

The results of this study have offered at least four important insights into the nature of language use on the internet. First, this study has demonstrated that the majority of internet texts can be reliably classified into web registers by non-expert internet users. During the course of this study our register framework evolved from a simple list of possible internet registers to a computer-adaptive web register classification instrument. This process included 10 rounds of piloting in which more than 100 people participated at various stages in the coding of 1,625 random internet webpages. The results of the final 1,000 webpage analysis showed that most web texts can be reliably classified into a general register and sub-register group or classified as a frequent register hybrid.

The second insight gained from this study relates to the third research question. Our method has revealed a great deal of register variation on the internet. Thirty-five of our fifty-six sub-register categories were agreed upon for at least one text. However, these results have also shown that a relatively small number of general registers and sub-registers accounts for a large proportion of the texts on the internet. For example, over 87% of all of the texts that were agreed upon were classified into one of the four most frequent general registers (Narrative, Opinion, Description, and Discussion). Furthermore, it was especially surprising to find that more than half of all the texts were classified as a News report/blog, an Opinion blog, or a Question/answer forum.

The third insight is the reality of register hybrids. The existence of internet texts with the characteristics of more than one traditional register category is generally accepted in the web-as-corpus community. However, our study is the first attempt at using a bottom-up approach to empirically identify register hybrids in a large-scale study. Using results from multiple participants, we have identified important two-way and three-way register hybrid categories that occur repeatedly on the web. While a great deal of future research will be needed in order to fully understand register hybrids on the web, the method used in this study seems to be a viable approach.

Finally, the results of this study can help us to understand the unique nature of language use on the internet. One of the most important attributes of internet texts is the potential for interactivity among multiple participants, often in real time. A particularly interesting finding was the fact that more than a quarter of all the webpages in our 1,000 URL contained reader comments. In a qualitative investigation of several sets of reader comments, we found that these comments are often initially written in response to the article. However, they soon transform into interactive discussions or debates among participants.

This study reports on the early stages of an ongoing, grant-funded project. In the next stage, we will use the web register classification survey in order to collect data on 50,000 random webpages. Once the texts have been classified into register categories, we will complete comprehensive linguistic descriptions of all the documents in the corpus. We will then evaluate the descriptive adequacy of the linguistic analysis, by determining whether the results of the linguistic analysis can be used to accurately predict the register of a web text. After possible revisions, we will automatically apply the register framework to a 100 million word web corpus. This corpus will be made freely available in its tagged and register-annotated form through Mark Davies' web-based corpus interface. The results of this study have the potential to inform our understanding of linguistic variability on the internet and improve NLP applications.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1147581.We also thank Anna Gates and Rahel Oppliger for their help with webpage classification.

References

Baroni, M and S. Bernardini (2004). BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004, Lisbon: ELDA. 1313-1316.

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

Biber, D. (1988). *Variation across speech and writing.* Cambridge: Cambridge University Press.

Biber, D. & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Crowston, S. (2010). Problems in the use-centered development of a taxonomy of web genres. In Mehler, A., Sharoff, S., & Santini, M., (eds.), *Genres on the web: Computational models and empirical studies*. New York: Springer.

Rehm, G., Santini, M., Mehler, A., Braslavski, P., Gleim, R., Stubbe, A., Symonenko, S., Tavosanis, M. & Vidulin, V. (2008). Towards a reference corpus of web genres for the evaluation of genre identification systems. In *Proceedings of the 6th Language Resources and Evaluation Conference*, pages 351-358, Marrakech, Morocco.

Rosso, M.A., & Haas, S.W. (2010). Identification of web genres by user warrant. In Mehler, A., Sharoff, S., & Santini, M., (eds.), *Genres on the Web: Computational Models and Empirical Studies*. New York: Springer.

Santini, M. (2007). Characterizing genres of web pages: genre hybridism and individualization. In *Proceedings of the 40th Hawaii International Conference on System Sciences (HICSS-40)*. Hawaii.

Santini, M. (2008). Zero, single, or multi? Genre of web pages through the users' perspective. *Information Processing and Management*, 44, 702-737.

Santini, M. and S. Sharoff. (2009). Web genre benchmark under construction. *Journal for Language Technology and Computational Linguistics*, 25(1):125–141.

Sharoff, S. (2005). Creating general-purpose corpora using automated search engine queries. In M. Baroni and S. Bernardini, (eds.), *WaCky! Working papers on the Web as Corpus.* Gedit, Bologna.

Sharoff, S. (2006). Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4), 435-462.

Sharoff, S., Wu, Z. K. and Markert. (2010). The Web library of Babel: evaluating genre collections. In *Proceedings of the Seventh Language Resources and Evaluation Conference, LREC* 2010, Malta.

Vidulin, V., Luštrek, M., & Gams, M. (2009). Multi-label approaches to web genre identification. *Journal for Language Technology and Computational Linguistics*, 24(1), 97-114.

Big and diverse is beautiful: A large corpus of Russian to study linguistic variation

Alexander Piperski[1], Vladimir Belikov[1], Nikolay Kopylov[2], Eugene Morozov[2], Vladimir Selegey[1, 2], Serge Sharoff[3]
[1]Russian State University for the Humanities, Russia
[2]ABBYY, Russia
[3]University of Leeds, UK
apiperski@gmail.com, vibelikov@gmail.com,
Nikolay_Ko@abbyy.com, Eugene_M@abbyy.com
Vladimir_S@abbyy.com, s.sharoff@leeds.ac.uk

Abstract

The General Internet Corpus of Russian (GICR) is aimed at studying linguistic variation in present-day Russian available on the Web. In addition to traditional morphosyntactic annotation, the corpus will be richly annotated with metadata aimed at sociolinguistic research of language variation, including regional, gender, age, and genre variation. The sources of metadata include explicit information available about the author in his/her profile, information coming from IP or URL, as well as machine learning from textual features.

1 Russian corpora: an overview

The linguists studying Russian have a wide range of different corpora available. By far the most popular resource is the Russian National Corpus,[1] which has become a de facto standard for the majority of corpus-based studies in Russian linguistics. However, this corpus is not well-suited for exploring the present-day language, since recently produced texts constitute a small proportion of it, and they are selected from a small number of sources. Other Russian corpora, such as I-RU, an Internet snapshot of Russian (Sharoff 2006) or ruTenTen[2] lack metadata. They are also often too small to capture frequencies of linguistic phenomena specific only to some part of the Russian-speaking community. Therefore many linguists have to rely on statistical data provided by search engines, such as Google or Yandex (the most popular search engine in Russia), but the drawbacks of this method are well-known (Kilgarriff 2007, Belikov et al. 2012).

2 General Internet Corpus of Russian: aims and objectives

The lack of a corpus representing the modern usage of Russian with diverse metadata gave rise to the General Internet Corpus of Russian (GICR) project which has been under development at the Russian State University for the Humanities since 2012 (cf. Belikov et al. 2012, Belikov et al. 2013). The aim of creating GICR is to provide the linguistic community with a reliable tool for studying the present-day Russian with specific information on language variation. In order to achieve this, it is necessary to collect a large amount of texts from the Web. The final version of the corpus is estimated to contain around 100 billion words by 2014.

The texts in GICR will be extensively annotated. Apart from morphological and syntactic annotation, GICR will contain a lot of metadata pertaining to the texts included in the corpus, such as gender, age, social status of the author, genre, topic and regional variety.

One specific objective is to draw attention to regional variation in Russian. It has always been acknowledged that there are village dialects in Russia (cf. Kasatkin 2005), but until recently the common opinion was that Russian of the cities is more or less homogeneous. However, this was questioned by Belikov (2006). His online dictionary *The Languages of Russian Cities*[3] shows that there are remarkable differences which span a wide range of uses, including locally produced legal texts (*vybit' chek* vs. *otbit' chek* 'issue a receipt'), professional terminology (*obnalichka* vs. *opanelka* 'door frame'), names of games or classes for schoolchildren, etc. Slight differences in morphosyntax are also existent, but a large

[1] http://ruscorpora.ru/en/
[2] http://trac.sketchengine.co.uk/wiki/Corpora/TenTen
[3] http://community.lingvo.ru/goroda/dictionary.asp

corpus with enough metadata is needed to investigate this issue.

3 Data collection and indexing

For data collection we use an adapted version of Nutch to crawl the Internet starting from the known hotspots of the Russian Web. The segments which are being investigated are blog platforms, forums, magazines and newspapers, etc. As of June 2013, GICR includes:

- the Russian-language blogs from LiveJournal.com, which is the most popular blog platform in Russia;
- the magazines form the *Magazine Reading Room* (*Zhurnal'nyj zal*, http://magazines.russ.ru/), a large online collection of Russian fiction magazines;
- the travel forum *Vinsky forum* (*Forum Vinskogo*, http://forum.awd.ru/)

Since we aim to study present-day Russian, only texts that are less than 5 years old are included in the corpus. Further on, we plan to adhere to this policy to keep GICR up-to-date.

At present, the size of the corpus is 1.38 billion words. However, it can be easily expanded, because making the corpus larger involves only a minimal amount of manual work.

Using blogs and forums, we expect to get most efficient results by keeping user profile statistics (date of registration, number of messages) together with user messages, so that we can get more benefits in analyzing site-specific user activity. Our algorithms rely on the idea that the more texts of a specific user we take into consideration, the more reliable the results of spam detection, age, and gender classification are.

Boilerplate removal algorithm is based on whether or not we know the web page structure (cf. Gibson et al. 2011). For known pages created using a well-known blog platform, content management system or forum platform, we can get only texts from the DOM element with well-known XPATH signatures. This also helps to separate the message body from the comments. For other pages we aim to employ a mixed strategy of taking the biggest contiguous block of text (Pomikálek 2011) or to use site-level boilerplate removal algorithms.

The page crawling strategy assumes that we collect all available web pages without using any page ranking function, but we only keep the content of those pages which have been created for humans, not for search engines. We put precision before recall, since the Russian Internet currently contains over 100 times more text than we plan for GICR.

The existing web interfaces, like Intellitext (Wilson et al. 2010) based on IMS Corpus Workbench (Christ 1994), or Manatee (Rychlý 2007) do not scale well to large corpora. Therefore we opted for development of a new system, using POS and shallow syntax annotated corpus in plain XML files indexed using what we can call *Narrowing Index*. Each sentence can be represented with some relatively small number, calculated as a product of prime numbers representing text features. Primes are assigned to word forms, lemmas, parts of regular expressions, and frequent bigrams of lemmas and RE in the descending order of their frequency. Each character number is connected with block number, by which we can reference the physical block in plain XML corpora. When we need to test a condition, the first step preceding plain corpus scan is finding block numbers in text corpora which may meet the condition of query. Random queries on an SSD storage are very fast, so that selective block retrieval becomes reasonably fast with a relatively small index.

An important type of queries concerns grouping the results, e.g. the `group` command in the CorpusWorkbench for producing collocations. Pre-caching of search results is not efficient because we cannot do set-theory arithmetic on partial results of sub-queries, but our users can be satisfied with the relative frequencies of the studied phenomena. We will collect partial results as soon as the frequencies converge to their practical limits. Since the Narrowing Index supplies us with a constant number of blocks where the query conditions are presumably met, the grouping queries can perform in constant time.

4 Text representation

The texts included in GICR are supplied with morphosyntactic annotation as well as metadata. We collected the web pages themselves (posts and comments are treated separately) as well as the following data from the user profile where available:

- username;
- user-chosen identification name (often identical with real name);
- year of birth;
- gender;

- region (this was unified to a standard form, also generalized to the respective administrative region)

Some of the authors provide only some part of this data. However, we had sufficient amount of training data even from this subset.

5 Text annotation

5.1 General issues

The size of the corpus implies that no manual annotation is possible, and for this reason it is crucial to choose fast and reliable automated annotation strategies. It is important to note that absolute accuracy cannot be achieved using such methods, but it is not a problem as long as corpus users are aware of this deficiency.

5.2 Morphosyntactic annotation

For morphosyntactic annotation we use an adapted version of the pipeline by Sharoff & Nivre (2011), which uses more Web-specific examples for training the POS tagger and the parser. The lexicon, especially for proper nouns and abbreviations, will be enriched as well.

5.3 Metadata annotation: processing pipeline

The starting point for metadata annotation is the explicit information about the author available in a standardized form in the profile of many blogging and forum platforms. Some information can be extracted from the IP address (server location for region determination) and URL (helpful for genre classification). All metadata of this kind are partial (not all bloggers provide it, IP addresses can be misleading, etc.), but this gives a source for training machine learning using textual features available on the page.

Text classification was based on standard extraction of lexical and POS features which provide sufficient reliability for this process (Sharoff et al. 2010), selection of keywords using the log-likelihood ratio (Rayson and Garside 2000) and using logistic regression and SVM for training. Because of the large amount of (sparse) training data, the Liblinear package (Fan et al. 2008) was used.

The dataset contains some amount of noise, which primarily includes spam pages (often automatically generated for search engine optimization), catalogues and other lists of objects, poems (which have very unusual text structure and linguistic properties). The majority of such instances have been cleaned by detecting the outliers using the values beyond 1.5 * *IQR* where *IQR* is the inter-quartile range for the following simple indicators:

- coverage by the most frequent words;
- average sentence length;
- text length in words.

5.4 Regional classification

There were two types of features used for machine learning. One comes from a specially compiled dictionary[4] which contains 710 words specific to different Russian-speaking regions. Other features were produced by selecting the keywords distinguishing each individual region from other regions using the standard log-likelihood keyness index (Rayson and Garside 2000). This procedure uses the top 800 words for each region (some words were specific for more than one region). For the preliminary classification, we selected 17 regions out of the complete set of Russian-speaking regions spanning over Russia and Ukraine. These regions are listed in Table 1. Moscow was excluded, since it is a melting pot for a large number of dialects.

Region	Country	Docs	%
Bashkortostan	Russia	53,420	4.29%
Chelyabinsk Oblast	Russia	49,798	4.00%
Donetsk Oblast	Ukraine	39,080	3.14%
Kiev	Ukraine	114,736	9.21%
Krasnodar Krai	Russia	50,544	4.06%
Krasnoyarsk Krai	Russia	41,032	3.29%
Moscow Oblast	Russia	119,328	9.58%
Novosibirsk Oblast	Russia	78,106	6.27%
Omsk Oblast	Russia	32,396	2.60%
Perm Krai	Russia	55,226	4.43%
St. Petersburg	Russia	300,814	24.15%
Rostov Oblast	Russia	64,340	5.17%
Samara Oblast	Russia	82,450	6.62%
Saratov Oblast	Russia	31,706	2.55%
Sverdlovsk Oblast	Russia	97,894	7.86%
Tatarstan	Russia	34,684	2.78%
	Total:	1,245,554	100%

Table 1: Regions and number of documents

For these regions, we used two sets of texts. One consisted of all texts longer than 20 words, the other included only texts longer than 300 words. The accuracy of regional classification (with 10-fold cross-validation) in the first case was about 15%, which is far from acceptable (the random baseline for the 17 regions would have been 6%). In the second case, it improved to

[4] http://community.lingvo.ru/goroda/dictionary.asp

35%, which shows that regional attribution for a very short text is very unlikely.

5.5 Gender classification

For gender classification, we used a collection of texts downloaded from the *Vinsky Forum* (http://forum.awd.ru). All posts by the same author were concatenated into a single text which was assigned the gender indicated by the author. All the words were lemmatized and supplied with grammatical features. The overall size of the collection using the format described above is 58,835 texts (28,200 texts by women and 30,635 texts by men). It is noteworthy that men tend to write more posts than women, because before concatenation we had 1,270,341 posts by men and 638,170 posts by women.

The best-suited machine learning algorithms for this purpose turned out to be logistic regression and SVM. Their results differed insignificantly, and only the results of logistic regression are provided here. All experiments included 5-fold cross-validation.

First, we tested POS-features such as the proportion of nouns, verbs, adjectives, pronouns, adverbs, prepositions, as well as the density of punctuation marks and the proportion of active voice verbs in a text. The results were low (precision = 0.574, recall = 0.575, F = 0.572). The average frequencies of different parts of speech are almost the same for men and women (the difference never exceeds 1%). This means that these features can hardly be helpful for gender classification.

Second, we chose three other features, namely the relative frequency of Adverb + Adverb bigrams (e.g., *very nicely*), of Adverb + Adjective bigrams (e.g., *very nice*) and of superlative adjectives. The accuracy of classification remained almost the same, but the number of features was significantly reduced.

Another approach was to use lexical classes described by Babych et al. (2007). The idea is to map words to general classes and to use the frequency of these classes as features. We excluded the lexical classes for which the frequency in male and female texts differed by less than 10%. The remaining classes with the correspondding male-to-female frequency ratios are represented on Graph 1. Swearwords also constituted a separate lexical class.

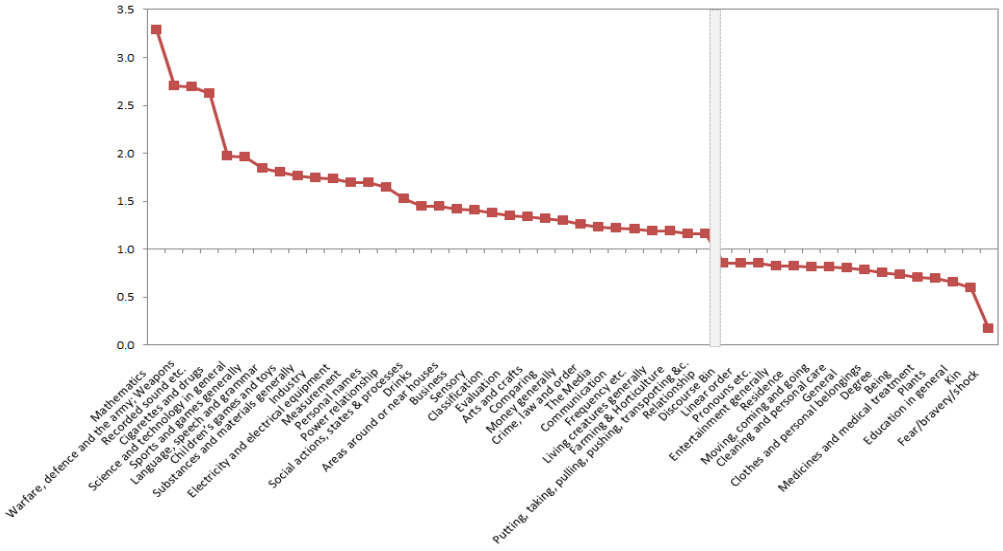
However, the problem is that some words belong to more than one lexical class, and these words had to be discarded. Unfortunately, such words are common among the most frequent ones. For example, *kuritsa* ‘chicken’ can belong to Food or Animals. A fast word-sense disambiguation algorithm would be useful for our purposes, but as long as such algorithms are still unavailable, we had to limit ourselves to a list of unambiguous words which contained 3000 items.

Further experiments combined lexical class features and POS features. Without imposing the lower limit on text length, we achieved the accuracy of 58%. However, such limits make it possible to improve the accuracy:

Number of words	Accuracy
≥ 1	58%
≥ 30	61%
≥ 200	67%
≥ 400	69%
≥ 1000	73%

Table 2: Accuracy of gender classification for texts of different length

There is one more grammatical feature of Russian that is useful for improving gender classification. Russian verbs are conjugated for gender in past tense (e.g., *ja*, *ty*, *on skazal* ‘I (masc.), you (masc.), he said’ vs. *ja*, *ty*, *ona skazala* ‘I



Graph 1. Male-to-female frequency ratio of different lexical classes

(fem.), you (fem.), she said'). For this reason, the bigram *ja* 'I' + past tense is highly indicative of gender. Of course, it may sometimes be misleading, but using this bigram as a feature for machine learning we were able to reach the accuracy of 77%.

6 Conclusions

GICR is a new corpus of Russian that will contain 100 billion words by 2014, which will make it a valuable resource for studying present-day Russian. Even now, it is already larger than the Russian National Corpus contains about 500 million words. GICR aims at raising awareness of sociolinguistic variation within Russian language, and the rich metadata in the corpus will provide a basis for studying this variation.

7 Acknowledgements

The present study was supported by the Minsitry of Education and Science of the Russian Federation and by the Program of Strategic Development of the Russian State University for the Humanities.

References

Babych, B., Hartley, A., Sharoff, S. & Mudraya, O. 2007. Assisting Translators in Indirect Lexical Transfer. In: Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), Prague, Czech Republic, June 23–30 2007.

Belikov, V. 2006. The examples for the dictionary of the varieties of urban Russian and the WWW. In *Proc. Computational Linguistics and Intelligent Technologies DIALOGUE2006*, Bekasovo, 57–60. (in Russian)

Belikov, V., Kopylov, N., Piperski, A., Selegey, V., Sharoff, S. 2013. Corpus as language: from scalability to studying variation. In *Proc. Computational Linguistics and Intelligent Technologies DIALOGUE2013*, Bekasovo, 84–96. (in Russian)

Belikov, V., Selegey, V., Sharoff, S. 2012. Preliminary considerations towards developing the General Internet Corpus of Russian. In *Proc. Computational Linguistics and Intelligent Technologies DIALOGUE2012*, Bekasovo, 37–50. (in Russian)

Christ, O., 1994. A modular and flexible architecture for an integrated corpus query system. In *Proc. COMPLEX'94*, Budapest.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Gibson, D., Punera, K., and Tomkins, A. 2005. The volume and evolution of web page templates. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, 830–839.

Kasatkin, L. 2005. Russian dialectology. Academia, Moscow. (in Russian)

Kilgarriff, A. 2007. Googleology is Bad Science. *Computational Linguistics* 33 (1): 147–151.

Pomikálek, J. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora*(PhD Thesis, Masaryk University, Brno, Czech Republic).

Rayson, P. and Garside, R. 2000. Comparing corpora using frequency profiling. In *Proc. of the Comparing Corpora Workshop at ACL 2000*, pages 1–6, Hong Kong.

Rychlý, P. 2007. Manatee/Bonito–a modular corpus manager. *Recent Advances in Slavonic Natural Language Processing (RASLAN)*, Masaryk University, Brno, 65–70.

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

Sharoff, S., Wu, Z., & Markert, K. 2010. The Web Library of Babel: evaluating genre collections. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, Malta.

Sharoff, S., & Nivre, J. 2011. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In *Proc. Computational Linguistics and Intelligent Technologies DIALOGUE2011*, Bekasovo, 591–604.

Wilson, J., Hartley, A., Sharoff, S., & Stephenson, P. (2010). Advanced corpus solutions for humanities researchers. In *Proceedings of the Pacific Asia Conference on Language, Information and Computation*. Sendai.

A Web Application for Filtering and Annotating Web Speech Data

David Lutz
Linguistics
Cornell University
del82@cornell.edu

Parry Cadwallader
Computer Science
Cornell University
wbc35@cornell.edu

Mats Rooth
Linguistics and CIS
Cornell University
mr249@cornell.edu

Abstract

A vast and growing amount of recorded speech is freely available on the web, including podcasts, radio broadcasts, and posts on media-sharing sites. However, finding specific words or phrases in online speech data remains a challenge for researchers, not least because transcripts of this data are often automatically-generated and imperfect. We have developed a web application, "ezra", that addresses this challenge by allowing non-expert and potentially remote annotators to filter and annotate speech data collected from the web and produce large, high-quality data sets suitable for speech research. We have used this application to filter and annotate thousands of speech tokens. Ezra is freely available on GitHub[1], and development continues.

1 Introduction

A vast and growing amount of recorded speech is freely available on the web, including podcasts, radio broadcasts, and posts on media-sharing sites. Much of this speech is accompanied by automatically-generated transcripts, and content providers and hosts often provide the ability to search these transcripts– and therefore the audio– for tokens of specific words or phrases. While these search features are usually designed for users to find content on topics that interest them, their potential use a source of speech data has not been lost on researchers. Howell and Rooth (2009) and Howell (2012) developed methods for automating data collection using these sorts of search engines using command line programs that interfaced with external search engines. This contribution continues the same research effort.

[1] https://github.com/del82/ezra

Whether collected automatically or not, the hits search engines return do not yet represent data for the linguistic or speech researcher. There remains a significant amount of work to convert these hits into useful speech tokens. First, each search hit must be manually filtered to determine whether it represents an actual token or is a false positive resulting from an error in the transcript. When the token is present, it must be extracted from the longer audio file it appears in, often with some surrounding context. Manual annotation is also frequently called for, depending on the nature of the specific study. For example, it may be necessary to record information about the speaker, the context, or other semantic, pragmatic, or discourse factors that may affect the token's acoustic properties or the way it was produced. The time required for researchers to filter and annotate hits in this way represents a major limiting factor in the efficiency of web speech data collection and therefore in the quantity and quality of web data that is available for speech research.

To address this efficiency challenge, we have developed a web application, which we call "ezra" that provides a simple but flexible interface for non-expert users to filter and annotate web-harvested speech data efficiently, and produce large and high-quality data sets suitable for speech research. Early versions of the application have been used within our own research group to process thousands of speech tokens. For one study, which examines examines the effects of semantic context on prosody, we collected tokens of the phrase *in my mind*. Published corpora were of limited utility for this study; transcripts of the Buckeye corpus (Pitt et al., 2007), for example, contain three tokens of this phrase. Using our application and freely-available audio from two radio stations,[2] we were able to collect and annotate

over 750 tokens of this phrase. By increasing the efficiency of filtering and annotating web audio, our application allows researchers to collect data to address very specific questions on a scale that had not previously been possible. The application is freely available.[3]

2 The application

Ezra fits into a workflow in which users collect speech data from the web containing potential tokens of the words, phrases, or other phenomena under investigation. Usually, the data are collected using the web harvest method detailed in (Howell and Rooth, 2009), where command line programs interface with a site that has indexed audio using automatic speech recognition (ASR). Once collected, data are imported into ezra, filtered, transcribed, and annotated based on the requirements of the the specific research being undertaken. Once processing is complete, the speech tokens, transcript, and annotations are exported for analysis.

Targets and hits

Annotation is centered around a *target*, which is a collection of search hits for a single word or phrase, like "in my mind", "some people", or "South Korea". Our specific research is concerned with targets that show focus prosody, or more generally have interesting patterns of prosody or variation in prosody. Each *hit* in a collection is a purported token of the target word or phrase. The goal of the application is to make as efficient as possible the process of *filtering* the hits, i.e. separating the genuine tokens of the phrase from the transcription and/or search errors, and *annotating* the genuine hits by correcting their transcripts and adding new information to their metadata.

Users

Users of the application are divided into two roles: supervisors and annotators. Supervisors are linguists who are working on a problem where a large sample of naturalistic uses of the target is expected

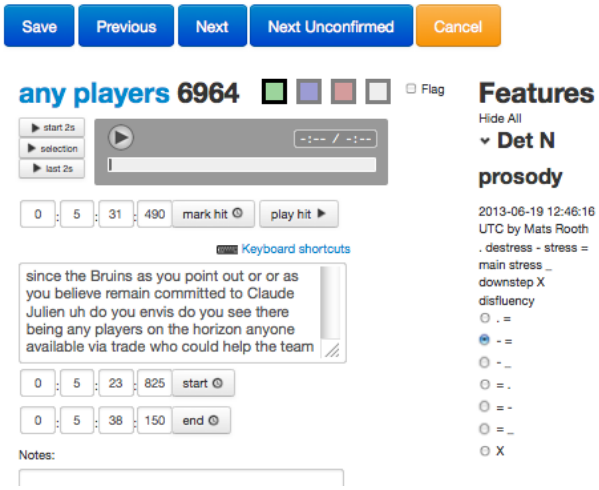


Figure 1: The result of annotating one hit. Boundaries have been marked, and the utterance transcribed. In the prosodic feature markup at the right, "-=" indicates a default prosody for "any players" where *player* is more prominent, but *any* still bears some stress. Buttons allow playing of the window, or of a shorter interval surrounding the target words. The Notes area at the bottom is used for free-form comments and interaction between users.

to provide evidence about theoretical issues, and to allow explicit models of the relation between acoustic form and linguistic levels (such as semantics and phonology) to be estimated.[4] Supervisors identify targets, arrange for data to be collected, import data, and export it from the application after a target has been filtered and annotated. Supervisors also design *features*, which are annotation tasks that are carried out for each hit of a target. Finally, supervisors are able to view the activities of other users, helping to monitor the progress of annotators and to allocate effort to different targets.

In contrast, annotators may not create targets or features, import or export data, or view other users' activities. Rather, annotators are focused on the filtering and annotation task. This division of roles, and attendant difference in privileges, allows annotators to focus on processing data and supervisors to concentrate on the tasks that precede and follow annotation.

[2]The radio stations were WNYC (http://www.wnyc.org) and WEEI (http://www.weei.com) which use (or have used) media search tools from RAMP (RAMP, 2011). http://www.ramp.com/case_study/weeientercom/ provides a case-study description of the RAMP audio search application at WEEI.

[3]Ezra is open-source and is available at https://github.com/del82/ezra. We welcome suggestions and contributions.

[4]See (Howell, 2012) and (Howell et al., 2013) for an example of this research program.

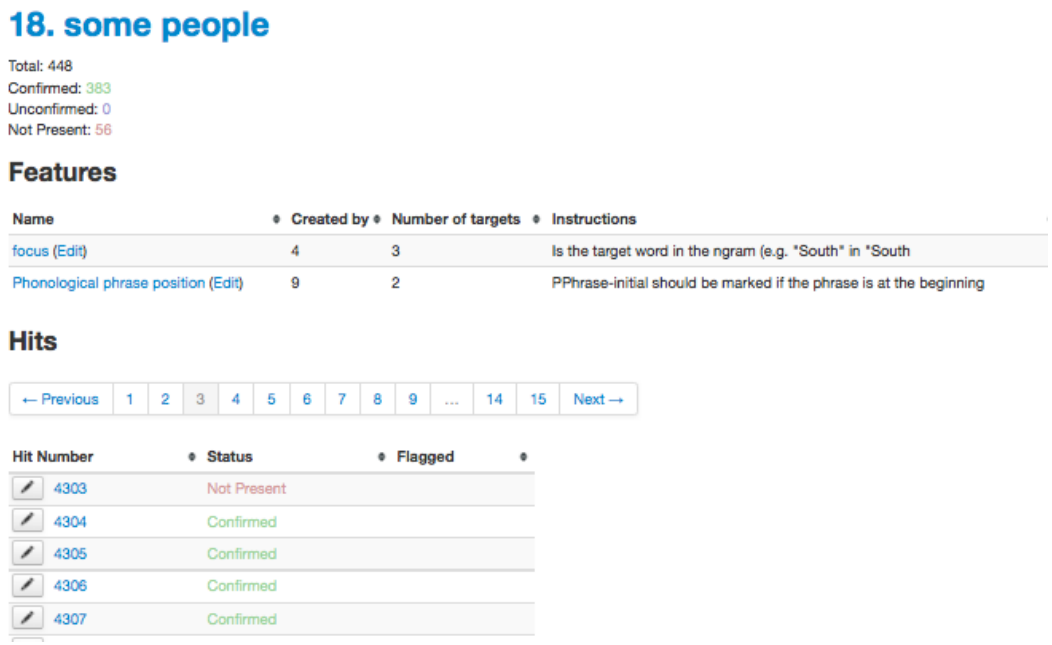


Figure 2: Part of the summary page for the target *some people*. Counts are at the top: of 448 hits, 383 were confirmed as containing the target, and 56 hits did not contain the target. The remainder were marked as repeats, or flagged as having problems of other kinds. The Features area summarizes the feature design. The Hits area at the bottom displays confirmation status of individual hits. Through links at the left, the annotation page for an individual page can be accessed, or a sequential record of annotation steps for the hit displayed, including the user who made the annotation.

Features

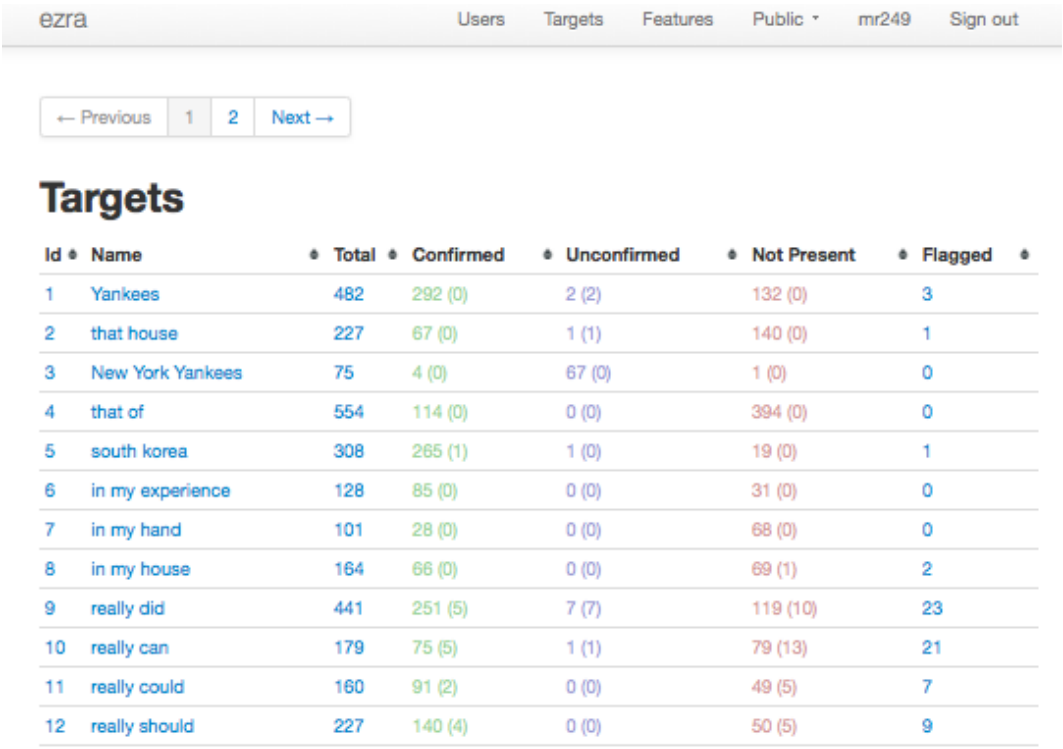
In addition to filtering the hits and correcting the transcripts surrounding genuine hits, supervisors may also specify other annotation tasks to be carried out when each hit is processed. These tasks, called *features*, are created within the web interface by a supervisor, and then assigned to a target. They may require the user to select one or more properties from a list, or they may ask for some text response. Features may include questions like: is the target focused or not? Was the token uttered by a man, woman, or child? Was it uttered by a native or non-native speaker? Each target may have multiple associated features, and each feature may be associated with multiple targets. See Figures 1 and 2 for examples. When processing each hit of a target, the user responds to all features that are assigned to it. The feature values are saved with the hit and exported along with the audio and the rest of the annotations.

The inclusion of features in ezra is designed to allow supervisors to include arbitrary annotation tasks with the filtering and transcription tasks. This ensures that each hit need only receive attention from a human once; after a hit is processed, its metadata will include all of the necessary information for the specific analysis for which it was produced.

3 Workflow

Once a target has been created in the system and hits have been imported, annotation begins. The annotator interacts with a hit through the web display seen in Figure 1. For each search hit, the annotator listens to the audio file around the time when it should, according to the ASR transcript, contain the target word or phrase. If the transcript is incorrect and the target is not present, the annotator notes that and moves on. Access to the ASR transcript, although it is imperfect, helps to orient the annotator and speeds up the filtering step. If the target is present in the audio, the annotator marks its exact location, and also marks the boundaries of a larger phrase or sentence in which the target appears, called the *window*. The annotator corrects the transcript of that window if necessary, and sets the values of each of the associated features for the token.

In our use of the application, for targets with no associated features but which required the annota-



Id	Name	Total	Confirmed	Unconfirmed	Not Present	Flagged
1	Yankees	482	292 (0)	2 (2)	132 (0)	3
2	that house	227	67 (0)	1 (1)	140 (0)	1
3	New York Yankees	75	4 (0)	67 (0)	1 (0)	0
4	that of	554	114 (0)	0 (0)	394 (0)	0
5	south korea	308	265 (1)	1 (0)	19 (0)	1
6	in my experience	128	85 (0)	0 (0)	31 (0)	0
7	in my hand	101	28 (0)	0 (0)	68 (0)	0
8	in my house	164	66 (0)	0 (0)	69 (1)	2
9	really did	441	251 (5)	7 (7)	119 (10)	23
10	really can	179	75 (5)	1 (1)	79 (13)	21
11	really could	160	91 (2)	0 (0)	49 (5)	7
12	really should	227	140 (4)	0 (0)	50 (5)	9

Figure 3: Targets page in ezra, giving summary counts for different targets.

tor to transcribe 10-15 seconds of audio surrounding the token, our annotators were able to filter and annotate about 60 hits per hour. Targets with more complex annotation requirements take longer; recent results suggest that about 45 hits per hour is achievable by an experienced annotator for these targets. Note that this rate depends on how many of the hits are false positives; a hit which does not contain a token is filtered in 20-30 seconds, but a hit which contains a token, and must therefore be annotated, may take three or four times that long to process. In our data, about 60% of the hits have been genuine tokens, while about 40% have been false positives or otherwise problematic.

When she creates each feature, the supervisor sets the possible values it may take for each token, and may include instructions for annotators, which serve as a ready reference while the annotator works. If an annotator is uncertain about how a particular token should be annotated, he may flag that token for further attention from the supervisor. Thus annotators can work without direct, immediate supervision without being forced to make decisions about which they are unsure, potentially introducing errors into the annotation.

Because ezra is a web application, annotators can work from anywhere, needing only a reasonably fast internet connection and a modern web browser. Robust user authentication and authorization allows the application to be deployed on the open web. Members of our group include researchers at three universities in two countries, and the application provides a shared environment in which to filter and annotate web speech data. Users may log in from anywhere, and only logged-in users may access the data. We found that the shared environment was important in coordinating our work. For instance, research leads for different targets can access ezra to check the progress of annotators, communicate about criteria for feature markup, and allocate annotation effort. Figures 2 and 3 show ezra pages that are used for examining summary results and the progress of annotation for a single target, and for all the targets together.

After the speech data has been filtered and annotated, the results are downloaded by the research lead for analysis. The download contains audio snippets, accurate transcripts, and the values of any features that were associated with the target. Each hit in the system is assigned a unique identification number on import, and retains this identifier through filtering, annotation, and export, allowing every audio clip to be uniquely identified and referenced in research and publications. In our

current workflow, after export we use the McGill ProsodyLab Aligner to create a phone-level alignment between the clip and a phonemic counterpart of the transcript. See (Gorman et al., 2011) and (Howell, 2012) for this methodology. Figure 4 shows a web presentation of the *any players* dataset that displays the alignment and allows audio to be played.[5] The alignment for hit 6964 is accurate. We found that in getting a good alignment, it is crucial to have a transcript that includes disfluencies (such as *uh* in Figure 4) and repeated words (such as *or or* in Figure 1).

In addition to the standard workflow, we have experimented with a pre-filtering workflow, where the research lead filters the data and marks approximate temporal boundaries. Then the annotator the creates the transcription and adjusts time boundaries to agree with word boundaries. This workflow has the advantage of allowing the lead to select on a theoretically-informed basis a window that includes the information which is relevant to what is going on in the discourse. For instance, for investigations of contrastive prosody, the preceding context may include an overt contrasting phrase that should be included in the window. While annotators working in the standard workflow also select a window which allows a listener to figure out what is happening in the discourse, the pre-filtering workflow allows the research lead to make the decision in a way that will allow specific hypotheses to be evaluated. The *any players* dataset seen in Figures 1 and 4 was created using the pre-filtering workflow.

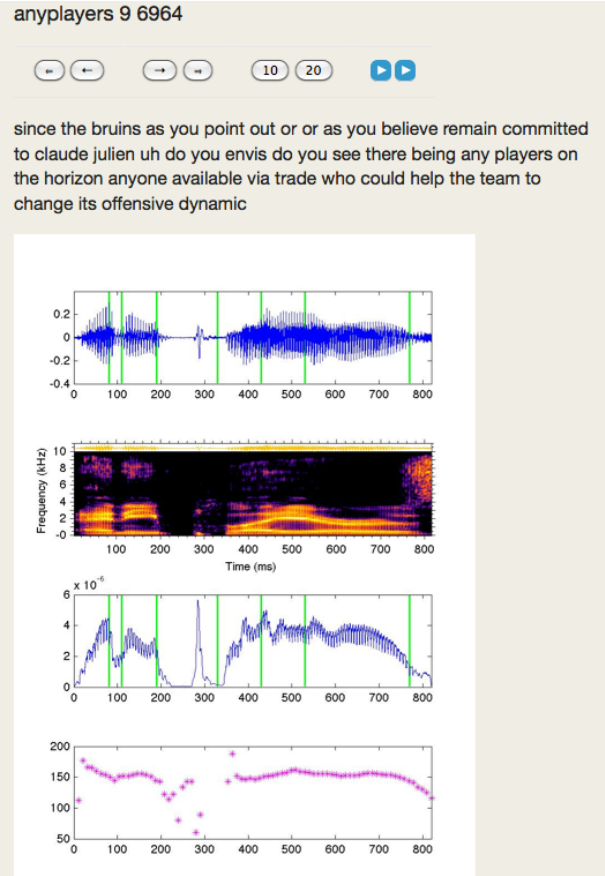


Figure 4: A web presentation of the *any players* dataset. The graphical panel displays an oscillogram, spectrogram, and plots of intensity and pitch. The green vertical lines indicate a temporal alignment for the seven phones of *any players* that was generated by using ProsodyLab Aligner.

Data collection and import

Before speech data can be processed in ezra, it must first be collected from the web. The specific type of data collected, and therefore its method of collection, depends on the goals of the research being undertaken. Ezra does not do the data collection, though we hope to add that capability through plugins in the future, as discussed in section 5.

In our work on prosody, our targets have been short two- or three-word phrases, and we have collected our data from two radio stations in the northeastern United States using media search functionality available on the stations' websites. The searches of these websites were automated using tools similar to those reported in Howell and Rooth (2009) and Howell (2012), which are able to conduct searches for specific phrases automatically and retrieve the URL of the audio file, the location in the file where the token is purported to appear, and in some cases a transcript of the audio surrounding the purported hit. This information is loaded into our application.

Because our own data collection relies on the media search functionality of content creators, the data available to us is limited to that provided by this search functionality. However, there is nothing inherent in the application that limits users to these types of searches or this type of data. The minimum requirement for each hit imported

[5]This web presentation, which is independent of ezra, is available at `http://compling.cis.cornell.edu/digging/`. In addition to the authors, Lauren Garfinkle contributed to the *anyplayer* dataset as annotator, and Kyle Gorman, Michael Wagner, and Jonathan Howell contributed in the implementation and tuning of ProsodyLab Aligner. Underlying audio data is property of WEEI. The graphical panel was produced with the Matlab code available at `http://www.cs.cornell.edu/~mats/matlab/`.

into the system is that it provide an audio file (local or remote), and a time in that audio file in which the target is purported to appear. Any method of searching or collecting media that produces this information can be used for data collection. For example, researchers may have a collection of audio files with transcripts containing speaker metadata, part-of-speech information, or syntactic or other structural information. Any time this information is automatically-generated, using e.g. automated speaker identification tools, part-of-speech taggers, or parsers, it is liable to contain errors. Ezra is designed to make the identification of false positives and the correction of errors as efficient as possible.

Put another way, ezra is intended to improve the *precision* of searches of audio data by allowing human users to filter and annotate the data quickly, removing false positives. It does not improve the *recall* of these searches, i.e. removing false negatives, as it has access to only those hits that the user imports.

In the future, we hope to develop partnerships with content providers that will allow us to address the challenge of limited recall, making the data our application produces more useful to language researchers for whom range of content, linguistic variety, and search recall are important considerations. Content providers might be willing to work with us to develop a way to search their audio which sacrifices precision in favor of recall– perhaps by providing the top three or top five most probable transcriptions generated by the speech-to-text system– which our human annotators could then correct, possibly helping them improve their transcripts and recognition system.

Despite these challenges in data collection, we view our application in its present state as one that can be of great benefit to researchers working with web speech data.

4 System Architecture

Ezra is written using the popular open-source Ruby on Rails[6] web application framework. It comprises a browser-based user interface, an application layer running on a web server, and a relational database in which user and application data is stored. Users access the application via a web browser. The interface is built using standard web technologies, including HTML5 generated with Ruby's standard ERB[7] template system, jQuery[8], and Twitter's Bootstrap[9] JS and CSS. The audio is played in an embedded player using SoundManager 2[10], which uses HTML5 to play the audio in browsers that support it and falls back on Adobe Flash for browsers that don't. The audio player has been specifically designed for the filtering and annotation task, and provides controls for playing only the audio window, only the token, or only the first or last two seconds of the window. Our annotators have found that these controls help them set the audio window and annotate the hit as efficiently as possible.

The server with which the client communicates is a Ruby on Rails application, which handles requests, including user authentication and authorization, and interaction with the database. Rails is open-source, mature, popular, well-documented, and straightforward to install on modern operating systems. Because of its popularity it is well-supported by other web technologies, and free tools exist that make it straightforward to deploy on a production web server such as Apache or nginx. At present, deploying and administering ezra requires some knowledge of Ruby and Rails, but we hope to reduce or eliminate this requirement as development continues.

Database

Ezra stores its data in a SQLite[11] database. SQLite is a simple lightweight relational database management system that stores the database in a file on disk. Rails provides a simple and powerful Object-Relational mapper through which the database can be accessed, and which provides a layer of abstraction which makes it possible to use, and migrate between, more full-featured database systems like MySQL and PostgreSQL with minimal changes to the application code. While this flexibility allows users to deploy ezra with their existing database infrastructure, we do not anticipate that usage volume will ever reach a level where SQLite is not fully adequate for our needs.

In addition to user, configuration, and audit data, the database contains records for each target, hit, and feature in the system. To simplify

[6] http://rubyonrails.org/

[7] http://ruby-doc.org/stdlib-1.9.2/libdoc/erb/rdoc/ERB.html

[8] http://jquery.com/

[9] http://twitter.github.io/bootstrap/

[10] http://www.schillmania.com/projects/soundmanager2/

[11] https://www.sqlite.org

updating and auditing, no records are ever deleted from the database. Users who are no longer a part of the project can be disabled, invalidating their login credentials and preventing access to the private parts of the system. Their user records remain in the database, however, because they contain a record of the work the user has done.

Information contained in hit records includes the audio file the hit appears in and its location within that file, whether the hit has been confirmed to contain the token of interest (or found to be a false positive), and all annotations associated with that hit. The hit record also contains two annotator-selected time points within the file that together demarcate the audio window containing the hit. This window is usually between 8 and 20 seconds long, and its boundaries correspond to the transcript the annotator produces of the hit. That is, the audio window is that portion of the audio in which the words in the transcript are uttered.

The audio files themselves are not stored in the database, but on disk alongside the database file. Each hit record contains the filename of this audio file, so that the application can serve the audio file along with the rest of the hit data. Like all records, hit records are retained indefinitely, even when a human annotator indicates that the target token is not present in the audio (a false positive), or when a hit is found to be a duplicate of another hit in the database.

Database statistics

Our ezra deployment contains, at the time of this writing, 9908 hit records of 35 targets, of which 6307 have been processed. Of those, 3928 (about 62%) have been confirmed as genuine tokens of their respective targets and annotated, 1971 (about 31%) have been marked as false positives, and 408 (about 6%) have been found to be duplicate hits (wherein the audio token indicated in the hit is an exact copy of another hit) or otherwise problematic. These numbers continue to increase as more hits and targets are added to the database and as annotators process hits. The SQLite database file containing this information is about 6 megabytes on disk.

5 Future work

In addition to working to improve the quality and quantity of data available to import into ezra, we continue active development on the application itself, driven by the feedback and suggestions from the researchers and annotators who are using it. While our first priority is always to make the filtering and annotation process as efficient as possible, there are several features we hope to add or improve in the near future.

- More complete user auditing and statistics would make it easier for supervisors to interact with annotators. The system keeps track of every change made to a hit, including the user who made the change, but not including the specific changes that were made. We'd like to improve this auditing functionality, and also make useful statistics available about the filtering and annotation work being done.

- We would like to add plugin functionality to both the import and export ends of ezra. For importing, allowing users to add integration with existing search engines and other data sources would greatly improve the quantity and diversity of data available. For exporting, plugins could integrate with other tools, e.g. the ProsodyLab Aligner (Gorman et al., 2011), or other manipulation or analysis tools.

- Although access to the audio and annotation functionality requires authenticated users, the application also serves publicly-accessible pages, which can be used to post papers and descriptions of the research being conducted. We would like to expand this functionality, making it easier for members to update the public pages via the web interface, and also to make selected audio and annotations available via the public site.

- Occasionally, data collection will generate duplicate hits, where two hits indicate the exact same audio token. These are not always from the same audio file, so comparing metadata will not always prevent duplicates. We would like to develop a method of detecting duplicate hits from the audio, and flagging them for further human examination.

- Because ezra is a web application accessible from anywhere, annotation work could be crowdsourced using e.g. Amazon Mechanical Turk, which has been used successfully

for transcribing spoken language data (Marge et al., 2010) and gathering linguistic judgments (Sprouse, 2011), though its use is not uncontroversial (Fort et al., 2011). We'd like to explore how we could update authentication and authorization to allow researchers to crowdsource their annotation if they so choose.

Ezra is open-source software, and its development is hosted in a public GitHub repository at https://github.com/del82/ezra/. This repository hosts the code, a public issue tracker, and a wiki containing user documentation that is being developed concurrently with the application. We invite fellow speech researchers to use ezra and contribute to its development with code, issue reports, feature requests, and contributions to the wiki.

As it stands, our application makes web speech data more accessible to researchers by providing a browser-based interface for filtering and annotating search results based on imperfect transcripts. It emphasizes simplicity in its interface, efficiency in its use of human annotators' time, and flexibility in the definition of annotation tasks and in the location of researchers and annotators. Although development and the addition of new features continues, the application has already been used to filter and annotate thousands of speech tokens, and represents a meaningful step in making the vast quantities of speech data on the web much more accessible for speech research.

Acknowledgments

Neil Ashton wrote the documentation available on GitHub. Ross Kettleson, Michael Schramm, and Lauren Garfinkle have been very patient and helpful annotators as ezra has been developed; Lauren was the annotator for the *any player* and *some people* datasets. Anca Chereches is the supervisor for the *some people* dataset. Jonathan Howell and Michael Wagner assisted by providing requirements and supervising the annotation effort at McGill. Scott Schiller, the author of SoundManager 2, provided help with functionality for playing sound intervals. Our research was supported by grant NSF 1035151 RAPID: Harvesting Speech Datasets for Linguistic Research on the Web (Digging into Data Challenge). This work is a contribution to a collaborative project in which Wagner and Howell are partners. The work of the McGill team was supported by SSHRC Digging into Data Challenge Grant 869-2009-0004.

References

Karën Fort, Gilles Adda, and K Bretonnel Cohen. 2011. Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*, 37(2):413–420.

Kyle Gorman, Jonathan Howell, and Michael Wagner. 2011. ProsodyLab-Aligner: A tool for forced alignment of laboratory speech. In *Proceedings of Acoustics Week in Canada*, pages 4–5, Quebec City.

Jonathan Howell and Mats Rooth. 2009. Web Harvest of Minimal Intonational Pairs. In *Web as Corpus 5*.

Jonathan Howell, Mats Rooth, and Michael Wagner. 2013. Acoustic Classification of Focus: On the Web and In the Lab. Ms. Brock University, Cornell University, and McGill University

Jonathan Howell. 2012. *Meaning and Prosody: On the Web, in the Lab, and from the Theorist's Armchair*. Ph.D. thesis, Cornell University.

Matthew Marge, Satanjeev Banerjee, and Alexander I Rudnicky. 2010. Using the Amazon Mechanical Turk for transcription of spoken language. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5270–5273. IEEE.

M.A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier. 2007. Buckeye Corpus of Conversational Speech (2nd release).

RAMP 2011. Universal Search Re-defined. White paper. http://www.RAMP.com

Jon Sprouse. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior research methods*, 43(1):155–67, March.

Compiling a diverse web corpus for South Tyrolean German - STirWaC

Sarah Schulz
LT3, Language and Translation Technology Team
Ghent University, Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
sarah.schulz@ugent.be

Verena Lyding, Lionel Nicolas
Institute for Specialised Communication and Multilingualism
European Academy of Bolzano, Viale Druso 1, 39100 Bolzano, Italy
{verena.lyding; lionel.nicolas}@eurac.edu

Abstract

In this paper, we report on the creation of a web corpus for the variety of German spoken in South Tyrol. We hence provide an example for the compilation of a corpus for a language variety that has neighboring varieties and for which the content on the internet is both sparse and published under various top-level domains. We discuss how we tackled the task of finding a balance between data quantity and quality. Our aim was twofold: to create a web corpus diverse in terms of text types and highly representative of South Tyrolean German. We present our procedure for collecting relevant texts and an approach to enhance diversity by detecting and filling gaps in a corpus.

1 Introduction

Creating large and diverse corpora for a language variety, as opposed to its standard, can be a particularly challenging task due to the smaller amount of data and the less ample distribution of text types available. In addition, it can be difficult to reliably distinguish between text of a variety and its neighboring varieties (including the standard one) and attribute data accordingly. These issues become even more cumbersome when dealing with web corpus creation as the standard procedure usually relies on the assumption that a great amount of text can be collected by simply restricting the search to relevant country-code top-level domains[1]. This procedure requires that the content of the selected domain or domains will be mostly written in the language of interest, which is rarely the case for language varieties besides the standard. In this paper, we first describe how we extracted text for South Tyrolean German; a variety that is not restricted to a single top-level domain and is scattered over several domains that also include text from other varieties of German. Among other tools, this procedure strongly relies on BootCaT[2] by Baroni and Bernardini (2004), a suite of scripts facilitating the compilation of web-based corpora. We then introduce a second procedure to address the challenge of compiling a corpus that contains dissimilar texts as defined by Forsyth and Sharoff (2013). This procedure aims at improving the balance of the corpus by detecting less represented types of texts and collecting new documents accordingly.

The main contributions of this paper are:

1. To detail a procedure for compiling a web-based corpus of a language variety that is not restricted to one single top-level domain and is considerably less represented on the internet than the standard variety.

2. To draft a procedure for improving the balance of a corpus in terms of (dis)similarity

[1] Country-code top-level domains (ccTLD) are two letter codes that identify pages allocated to a certain country or territory, e.g. *it*, *de*, *fr* etc.

[2] Download under: http://bootcat.sslmit.unibo.it/?section=download, 08.08.2012.

between documents by systematically searching and filling gaps.

3. To describe and evaluate the resulting STirWaC, the largest ever-built web-corpus for South Tyrolean German.

The remainder of the paper is organized as follows. We start with a general overview in section 2 where we situate our approach with regards to related research and provide a general outline of the whole process. We continue by describing two separate steps of corpus building. Section 3 describes the generation of a base corpus by BootCaT harvesting that then is used in a subsequent step, described in section 4, that extends the corpus through crawling. In section 5, we present an approach for adressing the issue of balancedness in terms of text types contained in a corpus. We evaluate our results with respect to related works in section 6. Finally, we discuss future works and conclude in section 7.

2 Overview

2.1 Related work

Using the world wide Web has become an established option for quickly building corpora. The approach however includes the challenging task of filtering the data for retrieving relevant documents.

As indicated by Cook and Hirst (2012), Murphy and Stemle (2011), Baroni et al. (2009), and Roth (2012), a country-code top-level domain can be used as a filter for national varieties of a language. The underlying idea is that most of the content published in the main language of a certain country is subsumed under its related top-level domain. For example, in the WaCky project[3], this concept has been successfully applied to build the deWaC corpus of German from the two top-level domains *.de* and *.at* whereas the top-level domain *.it* has been used for the itWaC corpus of Italian, etc. However, when not dealing with the main variety of a language, this approach does not apply unless the region in which a certain variety is spoken has its own top-level domain and the domain contains enough content. South Tyrolean German, like many other language varieties, does not meet these two conditions and thus represents a case that is not covered by the state-of-the-art.

[3] http://wacky.sslmit.unibo.it/

2.2 Global overview of the method

The process hereafter described can be divided into three major steps.

- Harvesting a base corpus.
- Crawling a larger corpus.
- Expanding the coverage over less represented text types.

The first step (see section 3), aims at quality above quantity whereas the second step (see section 4) is devised to improve quantity. Combining these two steps in a sequential manner, i.e crawling from a base corpus carefully collected, is a standard state-of-the-art procedure for optimizing both quality and quantity. However, as we will discuss in section 5, such approach could be integrated with a method for determining less covered text types and searching for new texts accordingly.

3 Harvesting a base corpus

We start the compilation of the STirWaC corpus by creating a base corpus that aims at data quality, in terms of content written in South Tyrolean German, and not data quantity. We implement diversified strategies for data harvesting that are applied in two subsequent iteration cycles.

3.1 Initial iteration

The initial iteration aims at creating a core corpus that will serve as basis for compiling an ample list of South Tyrolean seed terms to be used in the second iteration. The first iteration combines two conservative approaches in parallel: (1) harvesting based on top-level domain and (2) harvesting based on specialized terms.

3.1.1 Top-level domain *.it*

Following the approach mentioned in section 2.1, we focus on the top-level domain *.it*, as it is expected to host the major part of documents in South Tyrolean German, given that South Tyrol is a region of Italy. In addition, it is the German variety with the strongest official status used in Italy. Hence, we are looking for pages containing high-frequent German words under this top-level domain to build corpus I.1a and thus take advantage of the fact that German pages appearing under the *.it* domain are most likely from South Tyrol.

In order to implement the search, the 100 most frequent German words, as listed in Meier (1978),

are chosen as seed terms. These words belong to different parts of speech: the first five words of this list are *der, die, und, in* and *den* but there are also nouns like *Jahr* or verbs like *haben*. Out of these 100 words 500 random tuples of length three are generated. The tuples are used as search terms for the Bing search queries performed by BootCaT. The queries are restricted to websites detected as German-speaking by the search engine and appearing under the top-level domain *.it*. Moreover, we use a black list to exclude pages that would contain irrelevant content or could only lead to false positives, such as youtube or Wikipedia[4].

We retrieve the first 50 results of each query and receive exactly $50 * 500 = 25,000$ urls. The fact that every query yields the maximum limit of 50 pages is in line with our expectations since we use high frequent German words, which should be traceable on almost every page in German. After cleaning this url list from duplicates, 15,572 urls remain. From this url list, BootCaT extracts the corpus I.1a composed of a total of 11,070 clean[5] documents containing 9,658,731 tokens.

3.1.2 Specialized South Tyrolean terms

The second conservative approach starts from a list of terms that are exclusive for South Tyrolean German, i.e. terms that are very unlikely to appear in texts written in other varieties of German. Therefore, texts containing such terms are very likely to be written in South Tyrolean German.

Along this idea, 43 typical South Tyrolean terms are manually collected as seed terms. Out of these 43 terms (also containing multi-word expressions) 500 tuples of length two are randomly generated and used for harvesting. We choose a length of two, instead of three, because pages containing combinations of three of these terms are very rare and thus likely to be just collections of typical South Tyrolean terms. For this pass, we also use the negative list from the previous pass and moreover exclude results from the top-level domain *.de* [6]. As before, we request the first 50 results of each query. As expected, the upper limit of 25,000 results is not reached since the used seed terms are rare words that are combined in a random fashion.

After cleaning this url list from duplicates, 10,420 urls remain and BootCaT extracts the corpus I.1b with a total of 3,990 clean documents containing 4,108,360 tokens.

3.1.3 Initial corpus

We combine the two separately collected corpora I.1a and I.1b into one corpus. After cleaning, the resulting corpus I.1 contains 14,869 documents and 13,442,536 tokens. The overlap between corpus I.1a and corpus I.1b is of only 191 documents. Since it represents only 1,3% of corpus I.1, the two methods described above show a great complementarity. We now explain how this conservatively built corpus allowed us to continue with fully-automated subsequent steps.

3.2 Second iteration

In the second iteration we augment the base corpus by applying again the BootCaT-based harvesting approach. However, we start by deriving seed terms from the previously generated corpus I.1.

3.2.1 Seed term extraction

In order to determine the seed term list, we extract a frequency list from corpus I.1 and compare it to the frequency list of a reference corpus for German. We choose deWaC as reference corpus since it has been compiled similarly, contains contemporary vocabulary and is very large. For every word of our corpus, we use its frequency in both corpora to compute a pointwise mutual information measure (henceforth MI) as described in Evert (2005). These values allow us to evaluate how characteristic of South Tyrolean are the occurrences of all words in our corpus with respect to the reference corpus. For this, we first filter the vocabulary list by two thresholds designed to compensate known issues of MI: we discard words shorter than three characters and words which occur less than three times in the specialized corpus. We then rank the vocabulary list according to their MI score and keep the top 1000 words.

Comparing this new seed terms list to the South Tyrolean one from the first iteration, 13 of the 43 initial seeds reappear in the new list. This outcome is reasonable as the manually compiled list (first list) is aiming for distinctiveness, i.e. terms should only be used in South Tyrolean German but not in other varieties. To the contrary, in this step, we aim at capturing differences in terms of relative frequencies[7] as the pointwise MI value would

[4] There is no South Tyrolean Wikipedia.

[5] Clean as the default configuration of BootCaT defines it.

[6] We exclude *.de* as a trial run showed that pages from *.de* are mostly false positives for this step, e.g. travel reports.

[7] E.g. a certain word used in both South Tyrolean German

show. Consequently not all our original seed terms appear among the newly top-ranked ones.

3.2.2 Harvesting based on seed terms

From the 1000 seed terms we randomly generate 5000 tuples of length two. For each seed pair we request 30 search results, without restricting the results by top-level domain, as the results do not indicate it to be necessary. We again use the black list mentioned above. We retrieve 103,896 unique urls (out of a maximum of 150,000) and, after cleaning, the resulting corpus I.2 contains 25,719 documents and a total of 39,405,480 tokens. The corpora built so far are still small when compared to the ones of Austrian and Swiss German by Roth (2012) that contain 200 to 300 million tokens.

4 Crawling

Whereas the previous step is focusing on quality, this one is devised to increase quantity.

We use the open-source web crawler Apache Nutch[8] for crawling new documents from a list of seed urls. So as to counterbalance the large amount of German, Austrian and Swiss pages included in the subcorpora from the previous step, we extract all urls from our two corpora (I.1 and I.2) but those from top-level domains *.de, .ch* and *.at*. This results in two separate url lists (based on corpus I.1 and corpus I.2) that contain sites from top-level domains *.com, .it, .net, .org, .eu* and *.info*. For both crawler runs, the default configuration is used. Nutch is initialised with a link depth of three from the root page and we generate and fetch a new segment containing the top-scoring 1000 pages as suggested in the manual. Moreover, we force Nutch to skip images, files with suffixes *mov, exe, zip* etc. and URLs with slash-delimited segments that repeat three or more times to break loops. It also skips urls containing *?, *, !, @, =*.

4.1 Crawling twice and merging

The first crawling job in Nutch is started with 14,245 seed urls from corpus I.1 After cleaning the resulting list of urls from duplicates, we keep a list of 135,285 urls. Processing this list with BootCaT results in corpus II.1 which contains 45,888 clean documents. The second crawling job is started with the 25,719 urls from corpus I.2 that have been reduced to 4,625 seed urls obtained by replacing all urls with a common path by the common path itself, i.e. from these URLs only the single shortest URL per site was kept. This allowed to start BootCaT with a list of 65,554 unique urls. The resulting corpus II.2 contains 23,336 clean documents and 22,170,902 tokens. We then combine corpora II.1 and II.2 into corpus II and corpora I and II into our final STirWaC corpus. After the removal of duplicates and near-duplicates, we keep a corpus of 86,749 documents and 82,262,840 tokens. This means that many of the documents that are found in the crawling step are duplicates or near-duplicates. This could indicate that the strategies we implement are exhaustive and the amount of documents in South Tyrolean on the Internet is rather small. The development of the size of all subcorpora and the final STirWaC corpus is subsumed in Table 1 whereas the development of the distribution by top-level domain over the different subcorpora can be inspected in Table 2. As we can observe, we succeed, in the crawling step, in restricting the domains to keep the number of German, Austrian and Swiss pages low. The amount of documents of the other domains are increased by several hundred percent.

5 Patching to increase diversity

Our STirWaC corpus covers those web documents that are among the most accessible when combining the state-of-the-art approaches of BootCaT harvesting and crawling. Since we do not discriminate text types in our procedure, it also represents the diversity of documents written in South Tyrolean as present on the Web with the related bias in terms of ratio of text types.

We now introduce an approach for patching the STirWaC corpus with documents not reached by the standard BootCaT harvesting and crawling. The method builds on the assumption that specialized seed term lists, specific to subcorpora of certain text types, can be used to explore previously missed parts of the Internet. A patch corpus could then be harvested based on targeted seed term lists that have to be derived from subcorpora of specific text types within our STirWaC corpus, but not from STirWaC as a whole.

In order to implement the proposed approach, we need to tackle three tasks. First, we need to classify the documents according to their text types, which are so far unknown. Second, we need to establish a mean to group the texts into subcor-

and standard German but more used in South Tyrolean.

[8] `http://nutch.apache.org/`, 09.08.2012.

	Corpus	**I.1a**	**I.1b**	**I.1**	**I.2**	**I**	**II.1**	**II.2**	**II**	**STirWaC**
	Method	*Harvesting*	*Harvesting*	I.1a ∪ I.1b	*Harvesting*	I.1 ∪ I.2	*Crawling*	*Crawling*	II.1 ∪ II.2	I ∪ II
Setup	Domains	.it	¬{.de}	-	all	-	I.1 \{.at, .ch}	I.2[a]\{.de, .at, .ch}	-	-
	Seeds	100 terms	42 terms	-	1,000 terms	-	14,245[b] URLs	4,625 URLs	-	-
	Search Tuples	500 of length 3	500 of length 2	-	5,000 of length 2	-	-	-	-	-
	Max Results/Query	50	50	-	30	-	-	-	-	-
	Upper Limit	25,000	25,000	15,060	150,000	40,588	-	-	69,224	103,425
Results	Unique URLs	15,572	10,420	14,930	103,896	39,813	135,285	65,554	64,892	88,651
	DeDuper-ed Docs	11,070	3,990	14,869	25,719	39,502	45,888	23,336	63,923	86,749
	Tokens	9,658,731	4,108,360	13,442,536	39,405,480	50,734,333	29,777,384	22,170,902	47,869,771	82,262,840

Table 1: Summary of the corpus.

Corpus / Domain	$I.1a$	$I.1b$	$I.1$	$I.2$	I	$II.1$	$II.2$	II	**STirWaC**
.it	11,070 (100.0%)	1,256 (31.48%)	12,149 (81.71%)	3,551 (13.81%)	15,099 (38.22%)	30,573 (66.63%)	4,027 (17.26%)	32,759 (51.25%)	36,561 (42.15%)
.de	-	-	-	10,544 (41.00%)	10,544 (26.70%)	723 (1.58%)	537 (2.30%)	1,171 (1.83%)	11,668 (13.45%)
.at	-	373 (9.35%)	373 (2.51%)	2,779 (10.81%)	3,090 (7.82%)	116 (0.25%)	145 (0.62%)	215 (0.34%)	3,283 (3.78%)
.ch	-	126 (3.16%)	125 (0.84%)	989 (3.85%)	1,102 (2.79%)	75 (0.16%)	30 (0.13%)	104 (0.16%)	1,204 (1.39%)
other	-	2,235 (56.02%)	2,222 (14.94%)	7,856 (30.55%)	9,667 (24.47%)	14,401 (31.38%)	18,597 (79.69%)	29,674 (46.42%)	34,033 (39.23%)
total	11,070	3,990	14,869	25,719	39,502	45,888	23,336	63,923	86,749

Table 2: Distribution of top-level domains of all subcorpora.

[a] From these URLs only the single shortest URL per site was kept.

[b] This should be 14,371 but our exclusion pattern was a tad too generous.

pora, in order to generate targeted seed terms lists to orientate new searches. Finally, we need to verify that BootCaT-harvesting based on these seed term lists will in fact enable us to retrieve documents of the same text types. Differently put, the third task is concerned with evaluating if this approach can be effectively combined with the approaches of BootCaT harvesting and crawling that we implemented earlier. In the following sections, we address the first and third question while leaving the second one to future work (see section 7).

5.1 Assessing corpus diversity and text types

Our approach relies on the method developed by Forsyth and Sharoff (2013). In this method, a limited set of texts has been manually evaluated on several linguistic aspects. For each text, several features are automatically generated from the content and used as coordinates of a vector. The coordinates are then reduced to two and can thus be plotted on a 2D map. The reduction of coordinates is performed by maximizing a clustering criterion that takes into account the manual linguistic evaluations performed. Thus, the coordinates are reduced so that texts that are close according to the manual linguistic evaluations[9] appear close to one another on the 2D map.

We use the trained tool for standard German[10] to plot the STirWaC corpus (see Figure 1). The plotting reveals that our corpus, according to the criteria in Forsyth and Sharoff (2013), is satisfyingly diverse. This observation is in line with our assumption that the corpus collected should reflect the diversity of documents present on the Web.

Figure 1 allows to spot less populated areas on the plot, which in turn are indicative for the types of documents our corpus is lacking. However, since the 2D plotting model does not explicitly tell the text type of a document through its position, we just know that documents plotted close to each other have similar text types. Therefore, any conclusion on concrete text types remains subject to interpretation. So as to as to explore this assumption further, we aim at finding new documents that can be plotted in the gappy areas.

5.2 Getting new documents

Following the described approach for assessing the type of documents we are lacking, we aim to

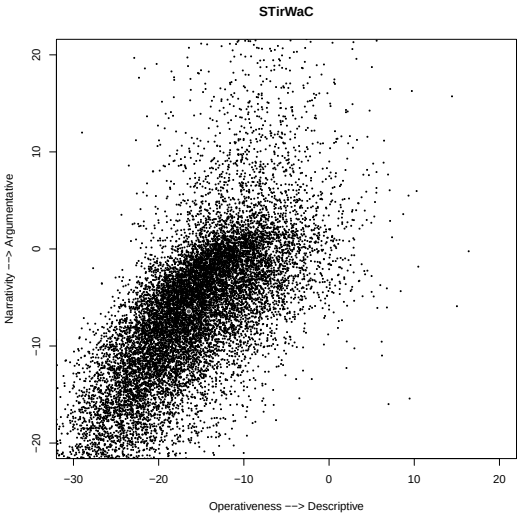


Figure 1: STirWaC corpus.

evaluate whether using this criterion to build subcorpora from less populated areas will indeed allow us to collect new documents fitting these gaps. In other words, we need to verify that this approach to extend the corpus not only by size but also by diversity of text types can be combined, in an effective manner, with the approaches for collecting documents that we implemented earlier.

Since we have not yet implemented a solution for grouping the texts into subcorpora, in order to generate targeted seed term lists to orientate new searches, we base our experiment on an external corpus instead of a dynamically compiled subcorpus. The main selection criterion for the external corpus is a ratio of text types as different as possible to the ratio in STirWaC, measured by plotting of the texts over less populated areas.

We chose to use the *Dolomitenkorpus* (DK), a collection of texts from a South Tyrolean newspaper[11] that are mostly not available online. As we see in Figure 2, the center of the plot of the DK is located at coordinates $[1.65, 2.39]$, that correspond to an empty area of the plot of STirWaC. According to our method, this indicates that the newspaper corpus contains text types that our STirWaC corpus lacks. Starting from this promising setup, a new seed term list is compiled from the DK as described in section 3.2. We compute the pointwise MI of each token by using STirWaC as refer-

[9] And are thus likely to be similar in terms of text types.

[10] Based on the above mentioned manual evaluations, cf. Forsyth and Sharoff (2013)

[11] Using the *Dolomitenkorpus* as seed term corpus for the initial compilation could have lead to a web corpus strongly biased towards newspaper texts, we thus originally excluded this corpus from our procedure.

ence. As the DK is larger than STirWaC and we want STirWaC to be the reference corpus, we take a sample of DK of 10% of the size of STirWaC. Such sampling allows to avoid non-specialized infrequent words that are present in DK but not in STirWaC. From the computed list of seed terms, we take the 1000 highest ranked words, create 5000 random tuples of length two and request the first 30 urls as result. From 150,000 potential urls we get back 67,784 unique urls of which 64,052 documents with a total of 75,225,045 tokens remain after cleaning.

As we can see in Figure 3, as expected, the plot of the newly harvested corpus, hereafter named *PatchCorpus* (PC), has a similar shape as the one for STirWaC and the center of its plot $[-12.54, -4.32]$ is located between the centers of the two plots of STirWaC $[-16.46, -6.43]$ and DK $[1.65, 2.39]$. Therefore, the center of PC is situated in a less populated area of STirWaC[12]. In addition, the angle between the vectors $(center_{STirWaC}, center_{PC})$ and $(center_{STirWaC}, center_{DK})$ is of 25.3 degrees. They thus have similar inclination[13] and we can conclude that using a corpus with a specific ratio of text types to compute seed terms allows to orientate the search and thus collect specialized texts. This indicates that the proposed approach can be combined with the approaches of BootCaT harvesting and crawling used so far.

6 Evaluation

A gold standard reference for evaluating a corpus of South Tyrolean German does not exist; therefore, we evaluate the corpus on the contained language rather than the diversity of text types.

To begin with, we used the chromium-compact-language-detector[14] to identify the language of each individual document, and 99,6% percent of the documents were identified as being written in German. This result was expected since we did discriminate on language when performing BootCaT harvesting and crawling, i.e. we explicitly searched for documents written in German.

[12]Having the center of PC located midway between the centers of DK and STirWaC and not completely overlapping the center of DK could be due to a possible bias implied by the search engine used.

[13]Perfect alignement in such 2D plotting where coordinates are not equivalent is unlikely.

[14]`https://code.google.com/p/chromium-compact-language-detector/` 20.06.2013

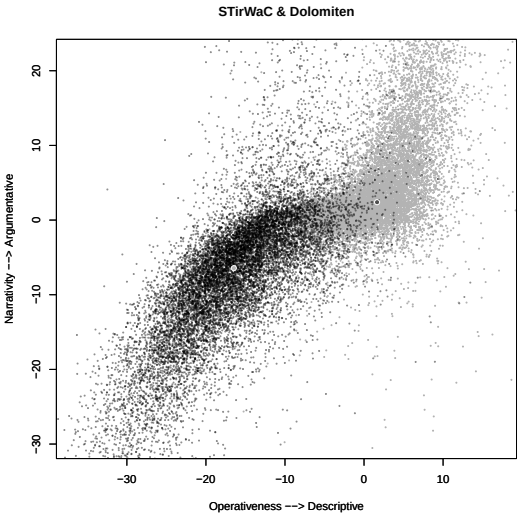


Figure 2: STirWaC corpus and Dolomitenkorpus.

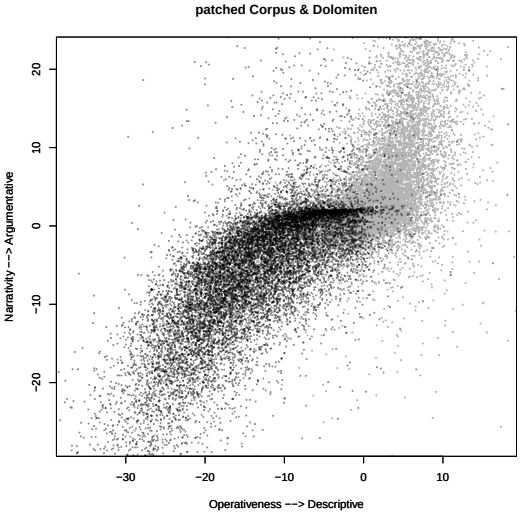


Figure 3: Patch corpus and Dolomitenkorpus.

collocation/term	Typical of	$\mathbf{rf}_{at}$	$\mathbf{rf}_{ch}$	$\mathbf{rf}_{de}$	$\mathbf{rf}_{st}$
wilder Knoblauch	*AT DE*	1.8	1.0	1.3	**4.9**
Blaulicht und Sirene	*CH DE*	2.2	**5.9**	3.7	2.4
Blaulicht und Folgetonhorn	*AT*	**4.0**	0	0	0
Blaulicht und Martinshorn	*DE*	1.8	1.5	**8.7**	0
in angetrunkenem Zustand	*CH DE*	0.7	**55.4**	2.0	37.7
Einspruch einlegen	*DE*	23.0	34.8	**90.8**	35.3
große Töne spucken	*DE*	**12.1**	9.8	11.8	0
Baukonzession	*STIR*	1.5	1.5	4.0	**305.1**
Handelsoberschule	*STIR*	0.4	0	0	**181.1**
Regionalrat	*STIR*	7.3	11.8	8.7	**494.8**
innerhalb <date>	*STIR*	0	0	0.3	**175.0**
halbmittag	*STIR*	0.4	0	0	**25.5**
weißer Stimmzettel	*STIR*	0	0	0	**6.1**

Table 3: Relative frequencies of characteristic n-grams over *STirWaC* ($\mathbf{rf}_{st}$) and three other corpora covering documents in Austrian German ($\mathbf{rf}_{at}$), Swiss German ($\mathbf{rf}_{ch}$) and the standard German ($\mathbf{rf}_{de}$) (Roth (2012))

Automatic language identification among different varieties of German is not yet available.

To check if documents of our STirWaC corpus globally belong to the South Tyrolean variety we therefore apply an approach similar to the evaluation performed by Roth (2012). The main idea of this evaluation is to compile a list of typical n-grams for each variety and manually check if their relative frequencies[15] in a corpus are representative for the language variety that this corpus is attributed to, i.e. if the relative frequencies of typical n-grams of a given variety var are high in corpora collecting documents written in the variety var and low in corpora collecting documents written in other varieties.

Additionally to the regionally marked multiword expressions (rows 1 to 7 in Table 3) suggested by Roth (2012), we randomly choose three words (rows 8 to 10) that have been marked as exclusively in use in South Tyrol in the 'Variantenwörterbuch des Deutschen' (Ammon et al., 2004) and that are not used as seed terms for BootCaT harvesting. In addition, we choose three typical collocations (rows 11 to 13) of South Tyrolean German described in Abel and Anstein (2011).

For all terms listed in Table 3, we compute relative frequencies over our STirWaC corpus and over representative corpora of the neighboring varieties of German: Austrian, Swiss and standard German. We aim to evaluate whether documents of the STirWaC corpus are more representative for South Tyrolean than for other German varieties.

The results in Table 3 show that relative frequencies of typical South Tyrolean terms (*STIR*) are significantly higher in our corpus than in the corpora of the other varieties. Also the frequencies of the terms that are characteristic for the other varieties provide confirmatory results. Indeed, relative frequencies of the n-grams typical for the other German varieties (*DE*, *AT* and *CH*) are low over our corpus. The numbers are comparable to the relative frequencies found in the other varieties that the n-grams are not characteristic for.

The frequency of the expressions evaluated by Roth (2012) fit into what would have been predicted. Just for the expression *in angetrunkenem Zustand* the frequency is higher than the frequency in Roth's German corpus, which contradicts the statement of Ammon et al. (2004) which predicts a higher frequency for this phrase for standard German than for South Tyrolean German.

[15]Normalized to occurrences per 100 million words following Roth (2012).

All together, these results can be taken as confirmation that our corpus is highly relevant for South Tyrolean German.

7 Conclusion and future work

This paper introduced *STirWaC* along with the approach that was implemented to build it. The current version of *STirWaC* contains 86,749 unique documents and a total of 82,262,840 tokens. It is the largest South Tyrolean web corpus currently available. The evaluation shows that it is highly relevant for South Tyrolean German.

We also presented the approach implemented to build *STirWaC*. This approach combines several state-of-the-art approaches and tools to harvest and crawl documents from the world wide Web. The practical results obtained confirm the relevance and validity of the presented approaches as well as the combination thereof. Consequently, we suggest that this approach can be used as blueprint for building corpora of other languages or language varieties; especially those for which the selection of relevant data from the Web in sufficiently large quantities is difficult.

Finally, we introduced a new approach to extend the state-of-the-art. This approach aims at extending both the size and the representativeness of a corpus by dividing it into subcorpora in order to devise specialized lists of seed terms for targeted new searches. Although this approach has not yet been fully implemented, we believe that the experiments we performed do demonstrate the relevance and viability of the underlying concept.

Future work is concerned with two main objectives: improving the size and the representativeness of the STirWaC corpus. Both objectives will be pursued by fully implementing the approach described in section 5. One important question left to tackle will be to determine how to select the subcorpora to build new, targeted seed terms lists. In that perspective, our current approach focuses on detecting gaps in the 2D projection and select the texts boarding them.

8 Acknowledgement

We would like thank Egon Stemle for recommending us the method of Forsyth and Sharoff (2013). We are also grateful to the reviewers for their useful comments and insights.

References

Abel, A. and Anstein, S. (2011). Korpus Südtirol - Varietätenlinguistische Untersuchungen. In *Korpusinstrumente in Lehre und Forschung*. University Press, Bozen.

Ammon, U., Kyvelos, R., and Nyffenegger, R. (2004). *Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol*. W. de Gruyter.

Baroni, M. and Bernardini, S. (2004). Bootcat: Bootstrapping corpora and terms from the web. In *LREC*. European Language Resources Association.

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

Cook, P. and Hirst, G. (2012). Do web corpora from top-level domains represent national varieties of english? In *Proceedings, 11th International Conference on Statistical Analysis of Textual Data / 11es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2012)*, pages 281–291, Liège.

Evert, S. (2005). *The statistics of word cooccurrences*. PhD thesis, Dissertation, Stuttgart University.

Forsyth, R. S. and Sharoff, S. (2013). Document dissimilarity within and across languages: A benchmarking study. *Literary and Linguistic Computing*.

Meier, H. (1978). *Deutsche Sprachstatistik*. Number Bd. 1 in Olms Paperbacks. Olms.

Murphy, B. and Stemle, E. W. (2011). PaddyWaC: A Minimally-Supervised Web-Corpus of Hiberno-English. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 22–29, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Roth, T. (2012). Using web corpora for the recognition of regional variation in standard german collocations. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*. Adam Kilgarriff and Serge Sharoff.

Web spam

Adam Kilgarriff
Lexical Computing Ltd. &
University of Leeds
UK
adam@lexmasterclass.com

Vít Suchomel
Lexical Computing Ltd. &
Masaryk University
Brno, Czech Republic
vit.suchomel@sketchengine.co.uk

Abstract

Web spam is getting worse. The biggest difference between our 2008 and 2012 corpora, both crawled in the same way, is web spam. In this paper we talk about what it is, with examples and a discussion of the overlap with 'legitimate' marketing material, and present some ideas about how we might identify it automatically in order to filter it out of our web corpora.

1 Introduction

Web spamming "refers to actions intended to mislead search engines into ranking some pages higher than they deserve" (Gyöngyi and Garcia-Molina, 2005). Web spam is a problem for web corpus builders because it is quite like the material we want to gather, but we do not want it. (We assume a 'general crawling' method for web corpus construction.)

Here are some examples:

> The particular Moroccan oil could very well moisturize dry skin handing it out an even make-up including easier different textures.
>
> Now on the web stores are very aggressive price smart so there genuinely isn't any very good cause to go way out of your way to get the presents (unless of course of program you procrastinated).
>
> Hemorrhoids sickliness is incorrect to be considered as a lethiferous malaise even though shut-ins are struck with calamitous tantrums of agonizing hazards, bulging soreness and irritating psoriasis.

It is on the increase: when we compare two corpora gathered using the same methods in 2008 and 2012, enTenTen08 and enTenTen12, the web spam in the later one is the most striking difference.

It is a moving target. The spammers and the search engines are in a game where the spammers invent new techniques, which will often work for a while until the search engines have worked out how to block them. Meanwhile the spammers will work out new techniques. The comments in this paper are likely to be of purely historical interest in the near future.

Our concern for web spam has been driven by specific corpus studies (all for English). In one we were investigating the term "Moroccan oil". In enTenTen08 it scarcely occurred, in enTenTen12 most occurrences were spam associated with the beauty products industry. In another we were investigating "on the ? store" and found that most instances for four of the top fillers for the variable slot, *web, net, internet, online*, were spam. In a third we were looking into rare words found in dictionaries, and checked in enTenTen12 for a word we did not know, *lethiferous*. Twelve of its fourteen instances in enTenTen12 were spam.

Most web corpus builders use a range of filtering strategies such as checking that documents have mostly common words, and a plausible proportion of grammar words: web spam that was not fairly similar to good text would largely be filtered out by these processes. The remaining web spam looks quite like good text.

1.1 Intermediate cases

Consider the text chunk below:

> MoroccanOil is an oil treatment for all hair types. Moroccan Oil is alcohol-free and has a patented weightless formula with no build up. Softens thick unmanageable hair and restores shine and softness to dull lifeless hair. Instantly absorbed into the hair. Moroccan Oil will help eliminate frizz, speeds up styling time by 40%, and provides long-term conditioning to all hair types. Are $20 shampoos and conditioners worth it? Can good hair-care products be found at the drugstore, or are the expensive salon products really superior? In this comprehensive guide to all things hair care,

Taken on its own this is respectable English. However there were many such pages, often with the same short sentences and sentence fragments in

different order or mixed in with less coherent and grammatical parts, often also on pages of "news items" with a 'read more' link at the end of each paragraph. The text is a marketing text, with component sentences written by a person, but that does not exclude it from being spam (on the definition we opened with). The line between marketing and spam is not easy to draw.

A recent development in this territory is 'content farms' where people are paid (poorly) for writing lots of articles, with the primary goal of driving traffic to advertising sites.[1] This is human-written and coherent, yet fits our definition of web spam. It is not clear whether we want it in a linguistic corpus.

2 Related work

(Gyöngyi and Garcia-Molina, 2005) present a useful taxonomy of web spam, and corresponding strategies used to make it. Their paper was presented at the first AIRWeb (Adversarial Inforrmation Retrieval on the Web) workshop: it was the first of five annual workshops, associated with two shared tasks or 'Web Spam Challenges'. The last of the AIRWeb workshops was 2009; in the years since, there have been joint WICOW/AIRWeb Workshops on Web Quality.[2] These workshops, held at WWW conferences, have been the main venue for IR work on web spam.

Since the merge, there has been less work on web spam, with the focus, insofar as it relates to spam, moving to spam in social networks and tagging systems (Erdélyi et al., 2012).

The datesets used for the shared tasks are called WEBSPAM-UK2006 and 2007 and are described in (Castillo et al., 2008). Labels (spam or non-spam) were at the level of the host rather than the web page. A large number of hosts were labelled in a substantial, collective labelling effort: 7473 hosts in UK2006 and 6,479 in UK2007. UK2006 had 26% spam whereas UK2007 had 6% spam: the difference is because UK2006 did not use uniform random sampling of a crawl whereas UK2007 did, so 6% is the useful figure for reference. The tagged data was split with two thirds usable for training, one third retained for evaluation. There were six participants for UK2007 and all used supervised machine learning, with a range of text-based and link-based features, and the best system scoring 85% 'area under curve'. This was improved upon by (Erdélyi et al., 2012), who also discuss the ECML/PKDD Discovery Challenge dataset where 'spam' is one of a number of labels.

2.1 Search Engines

Web spam is a game played between spammers and search engines. Search engines —particularly the market leader Google, also Bing, Yandex, Beidu— employ teams of analysts and programmers to combat spam. In those companies there will be great knowledge of it and expertise in identifying it. They probably have large recent databases of spam, to conduct experiments on. However these resources and expertise will not, for obvious reasons, be shared outside the company. A good feature of AIRWeb is that representation on it from search engine companies is high: Carlos Castillo, from Yahoo, notes in his powerpoint reviewing the Web Spam Challenges[3] "keeping web data flowing into univesities" as a goal and a benefit of the Web Spam Challenge.

The Google paper "Fighting Spam"[4] describes in broad terms the kinds of spam that Google finds, and what they do about it. Figure 1 shows developments from 2004 to 2012.

The BootCaT method for building corpora (Baroni and Bernardini, 2004) works by sending seed terms to a search engine, and gathering the pages found by the search engine. In this approach, the corpus-builder benefits directly from the search engine's measures against web spam.

2.2 Test data and evaluation

It is a big methodological challenge to gather a good sample of web spam. It is, by design, hard to find and set apart from good text. We can gather samples by simply noticing and putting spam documents to one side to build up a spam corpus. This is useful and probably central to all we might do, however in does not help us find the spam types we have not yet noticed.

Historical datasets are of limited value as spammers will have moved on: despite that, the WAC community will almost certainly benefit from using the AIRWeb and ECML/PKDD datasets discussed above, and the filtering methods developed

[1] http://readwrite.com/2010/11/17/content_farms_top_trends_of_2010

[2] WICOW stands for "Workshop on Information Credibility on the Web".

[3] http://airweb.cse.lehigh.edu/2009/slides/castillo-challenges.pdf

[4] http://www.google.com/insidesearch/howsearchworks/fighting-spam.html

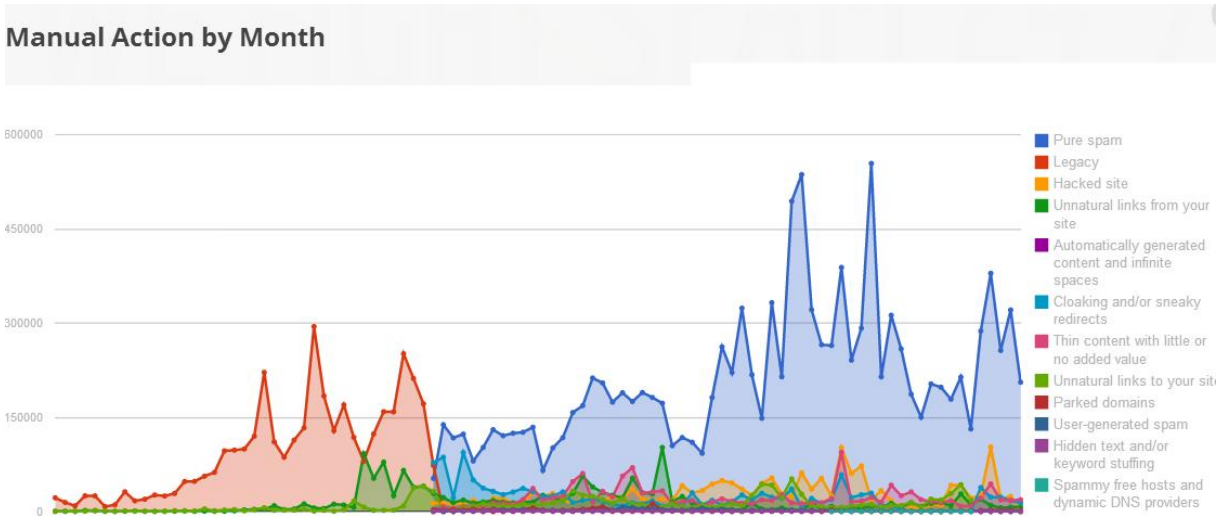


Figure 1: Google's analysis of spam types and quantities, 2004-2012.

there.

2.3 Level of analysis

The IR work mostly focuses on finding bad hosts (and much of it, on links, "the web as a graph"). That is a distinct strategy to finding bad text, e. g. within a web corpus once it has been cleaned, with links deleted. One question for web corpus builders is: at what stage should spam detection take place - before html-removal, or after, and do we work at the level of the page or the website? Also, should we concentrate on hosts, or domains, or web pages? Some prelimary evidence suggests that the landscape hosts and domains change very quickly, so methods based on text may retain validity for longer.

3 Methods

3.1 Coherence approaches

To recognise the examples above as web spam, we have to read them. This is in contrast to, for instance, noticing unwanted material not in English, or lists of English words, where a cursory glance is sufficient and the level of attention that deserves the word 'reading' is not required. The spam is not obviously grammatically flawed. But it lacks coherence. This suggests that, to identify it, we want to measure the coherence of each sentence or text, in order to identify spam as the low-scoring material.

Ways in which we might do this are:

- apply the entity-grid model of (Barzilay and Lapata, 2008);
- perform syntactic analysis to create dependency trees to model dependencies of parts of sentences. A "nice" tree could mean the sentence is coherent;
- in a coherent text we expect words to be from compatible domains and registers. It may be possible to identify sets of words that belong together (in terms of domain or register) and then to spot texts where words come from mismatched or incompatible domains or registers.

3.2 Words for things that people want to sell, and marketing buzzwords

Much web spam works to sell products, so the names of the things being sold will often be mentioned, as in Moroccan oil. Spammers will also use low-content terms that they think people will search such as "web store". If we can gather a long list of these items, we can use counts for them as part of a scoring system. (Baisa and Suchomel, 2012) explore this method, using a small spam corpus to identify n-grams which are notably more frequent there than in reference text.

3.3 Dictionary words

Lethiferous points to spammers using dictionaries to flesh out the lingusitic profile of their spam. Perhaps texts containing words which are in big, traditional dictionaries but have low corpus frequencies can act as alarm bells.

For removal of duplicates and near-duplicates in our corpora we use onion (Pomikálek, 2011). However we have recently noted that some web

spam avoids detection through random changing of content words to synonyms, drawn from a thesaurus. This is a method that could be reverse-engineered.

4 EnTenTen12 *vs.* EnTenTen08

We stated above that the biggest difference between EnTenTen12 and EnTenTen08 is web spam.

(Kilgarriff, 2012) presents a method for exploring differences between corpora, demonstrating how the manual classification of the top 100 keywords of corpus1 *vs.* corpus2 and vice versa gives a rich picture of the contrasts between the two. This is what we have done in this case, as follows:

- For each word matching
  - Find frequencies in corpus1 and corpus2
  - Normalise to 'per million'
  - Add a 'simplemaths parameter' of 0.001 to normalised figures (including the zeroes). This low value for the parameter means that the list will be dominated by low-frequency keywords.[5]
  - if the figure for corpus1 is larger than that for corpus2, divide the corpus1 figure by the corpus2 figure to give a score
- sort the words according to the scores

The highest-scoring words are the keywords for corpus1 *vs.* corpus.

One typically finds many names and nonwords in the lists so generated, and we were interested in dictionary words. We filtered to give only all-lower-case-letters items of length at least 3, and hunspell,[6] to give a list of words that were only 'dictionary words'. Table 1 shows the top of the list complete with frequency figures, to show the sheer magnitude of the differences in frequencies: 18,102 occurrences of *jewelries* in 2012 against 35 in 2008. Table 2 gives the full analysis.

Of the 100 words, six related to new things: three (*tweeting tweeted twitter*) to twitter, launched in 2006 with meteoric growth since 2007; *voltaic*, almost always in the context of photo voltaic cells, newly topical with climate change and associated government initiatives; *atomizer*, for which all the data related to electronic cigarettes (of which an atomizer is one part), which first appeared on the international market in 2005-06,[7]; and *jailbreak* which is what you do when you convert an Apple device such as an iPhone or iPod from one that can only operate in the Apple-approved ways to a general purpose device. In addition there was one new word, *colorway* (in both singular and plural; a synonym, widely used by clothing and footwear manufacturers, for *colour scheme*: "we have this design in all sizes and colorways") and *aftereffect*, increasingly spelt as one word.

Of these, *atomizer* and *colorway* relate to things that are marketed extensively on the web. So do most of the other 91 items. The straightforward shopping items are clocks and watches (six words), footwear (five), handbags and holdalls, birthstones (singular and plural), *pantyliners, jerseys, headpins* and *foodstuffs*. Services were financial (six items), locksmiths (two), *refacing* for kitchen cabinets and four words relating to weddings.

'Health and beauty' accounted for 28 of the 100 keywords. The leading subcategory is skin, with particular emphasis on spots. We have pimples, blackheads, whiteheads, moisturizers and dehydrators. The meaning of *breakouts* that put it in the keyword list was "a breakout of acne" and a *concealer* was always a concealer of acne.

There were just two items of a *lethiferous* flavour: *accouter* and *osculate*. *Accouter*, a rare synonym for *dress* (as in *accoutrements*) was widely used in spam associated with clothes and weddings. *Osculate*, a rare synonym for *kiss*, in spam associated with pornography.

The remaining large category was formed of words in morphological forms that were unusual for them: eleven nouns ending in '-ness', six plurals, two nouns and an adjective in -er, and two adjectives with -able.

The -ness nouns included *humorousness, severeness, comfortableness, anxiousness, courageousness neglectfulness, safeness*. These are odd because it is usual to use *humo(u)r, severity, comfort, anxiety, courage, neglect, safety* instead.

The plurals include mass nouns *attire, apparel, jewelry* which, in the first author's British dialect, scarcely bear pluralising at all.

The items *anticlimaxes, dejecting, unexceptionally* all have something contradictory about them. An anticlimax only exists in contrast to an ex-

[5] See (Kilgarriff, 2009) for discussion.

[6] http://hunspell.sourceforge.net/

[7] Wikipedia: *Electronic cigarette*

Word	enTenTen12		enTenTen08		Score
	Freq	Norm	Freq	Norm	
tweeted	28711	2.2	11	0.0	507.41
jewelries	18012	1.4	35	0.0	118.72
tweeting	26024	2.0	67	0.0	93.40
colorway	6395	0.5	17	0.0	79.69
hemorrhoid	57951	4.5	181	0.1	79.29
straighteners	28206	2.2	133	0.0	52.20
courageousness	8717	0.7	40	0.0	50.86
twitter	712447	54.9	3602	1.1	49.81
straightener	23324	1.8	137	0.0	41.94
colorways	4242	0.3	23	0.0	40.83
anticlimaxes	2584	0.2	14	0.0	37.91
wagerer	1060	0.1	4	0.0	37.21

Table 1: enTenTen12 top keywords, showing figures and working.

pected climax, and climaxes tend to be singular by their nature, so it is hard to see a role for the plural version of their contrasts. The verb *deject* is always passive so it is hard to see how something can be dejecting. *Exceptionally* brings attention to the predicate it is associated with: when we negate it with -un it is unclear what we are doing.

Discussion

All 100 words except the three twitter words and *voltaic* were highly associated with spam, as confirmed by scanning concordances. For some –*wagerer, osculate, conveyable*– all of a sample of fifty concordance lines appeared to be spam, but for the majority, the judgement was not easily made, with most of the sample being on the spectrum between marketing and gibberish.

For the shopping, services, and health-and-beauty words, we see the results of spammers taking legitimate material, chopping it into pieces and permuting and varying it.

The morphology cases are more puzzling. Three hypotheses for the radical increases in frequency of these terms are:

1. A computer is generating derived forms of words and using them in spam: example

   > This, in addendum to modern sedate safeness concerns, numberless increases in data sum total, and rising cost pressures, closest these organizations with some uncommonly outstanding topic challenges.

2. Authors are non-native speakers of English. They will often use the regular nominalisation (*anxiousness*) rather than the irregular one (*anxiety*) and pluralise mass nouns in error. The following seems likely to be a non-native production:

   > The minimum height I would suggest for your inside rabbit cage would be 40 cm, but this only a guide. Please use you discretion and if in doubt go for the taller cage. A lot of individuals choose for numerous floor bunny rabbit cages with brings joining the levels. This grants the bunny rabbit a lot extra room without borrowing more room inside your haven. Owning a line flooring inside your bunny rabbit Cage isn't a good plan if you would like to give **comfortableness** for your bunny rabbit. While having a wire bed with a pull out and makes for simpler maintaining, it's not all of the time necessary as bunnies are easily litter box trained.

3. It is a matter of dialect: whereas the first author will always say *comfort* rather than *comfortableness*, and for him, *jewelries* is close to impossible, this is not so in other dialects. (Kachru, 1990) discusses the varieties of English in terms of the *inner circle* (the traditional bases of English: UK, USA, Australia, New Zealand, Ireland, anglophone Canada), the *outer circle*: countries where English is historically important and is central to the nation's institutions; South Africa, India, Nigeria, the Philippines, Bangladesh, Pakistan, Malaysia, Kenya; and the *expanding circle*, where English is playing a growing role, which covers much of the rest of the world. The inner circle countries are all high-wage, so it would not be surprising if companies looked to outer-circle countries, where there are both many speakers of local dialects

```
NEW THINGS
  tweeting tweeted twitter
  (photo) voltaic (cells)
  atomizer (as part of aparatus for giving up smoking)
  jailbreak (verb: remove limitations on an Apple device)

NEW WORDS
  colorway colorways aftereffect (increasingly spelt as one word)

SHOPPING
  footwear espadrille sneaker slingback huarache
  handbags holdalls
  chronograph chronographs timepiece timepieces watchstrap watchmaking
  birthstone birthstones
  foodstuff
  headpins (jewelry making)
  pantyliner jerseys

SERVICES
  locksmith locksmiths refacing (for kitchen cabinets)

MONEY
  refinance refinancing remortgages defrayal cosigner loaners

WEDDINGS
  bridesmaid boutonnieres honeymoons groomsmen

HEALTH AND BEAUTY
  periodontist whitening veneers aligners (both mainly for teeth)
  hemorrhoid hemorrhoids
  hairstyles straightener straighteners
  slimming physique cellulite liposuction stretchmarks suntanning
  moisturize moisturizes moisturized dehydrators detoxing
  pimples whiteheads blackhead blackheads
    breakouts (of acne etc) concealer concealers (of acne etc)
  tinnitus

RARE DICTIONARY WORDS
  accouter osculate

MORPHOLOGY
  humorousness severeness sturdiness impecuniousness comfortableness
    anxiousness adorableness courageousness neglectfulness moldiness safeness
  anticlimaxes chitchats attires apparels jewelries jackpots
  wagerer vacationer dandier
  acquirable conveyable
  dejecting unexceptionally

NAMES (incorrectly included - most were filtered out)
  spellbinders (company) circuital (album) android (operating system)

OTHER
  frontward proficiently
```

Table 2: An analysis of the top 100 keywords of enTenTen12 *vs.* enTenTen08 (simplemaths parameter=0.001, filtered to give only all-lowercase dictionary words at least three characters long). All capitalised text is authors' lables for categories, and all text in brackets is explanatory glosses. All other words are the keywords.

of English, and low wages, to write bulk marketing material for SEO. Consider:

> It is dream of every woman to have a perfect wardrobe. The thing that tops the list to make the wardrobe a complete one is a black shoe. Ladies black shoes add style and versatility to the **attires**. From casuals to formal black is the colour that makes the feet stand out from the crowd.

To the first author's British ear, this sounds like Indian English.

5 In sum

Web spam is a large and growing problem for web corpus builders, at least for English. There has been work on it in the IR community (to date, to the best of my knowledge, not known to the WAC community). The WAC community can benefit from that work.

We have also presented some linguistic observations that could prove useful for spam identification, and some data relating to changes we have observed between 2008 and 2012.

6 Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013. The access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005) is highly appreciated.

References

Vít Baisa and Vít Suchomel. 2012. Detecting spam content in web corpora. In *Recent Advances in Slavonic Natural Language Processing (RASLAN-6)*, Masaryk University, Brno, Czech Republic.

Marco Baroni and Silvia Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *In Proceedings of LREC 2004*, pages 1313–1316.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

C. Castillo, K. Chellapilla, and L. Denoyer. 2008. Web spam challenge 2008. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*.

Miklós Erdélyi, András Grazo, and András A. Benczúr. 2012. Web spam classification: a few features worth more. In *Proc. Joint WICOW/AIRWeb Workshop at WWW-2012*.

Zoltán Gyöngyi and Hector Garcia-Molina. 2005. Web spam taxonomy. In *AIRWeb, Proceedings of a Workshop on Adversarial Information Retrieval on the Web*, pages 39–47.

Braj Kachru. 1990. *The alchemy of English: the spread, functions, and models of non-native Englishes*. University of Illinois Press.

Adam Kilgarriff. 2009. Simple maths for keywords. In *Proc. Int. Conf. Corpus Linguistics*, Liverpool.

Adam Kilgarriff. 2012. Getting to know your corpus. In *Proc. Int. Conf. Text, Speech, Dialogue*, Brno, Czech Republic.

Jan Pomikálek. 2011. *Removing Duplicate and Boilerplate Content from Web Corpora*. Ph.D. thesis, Masaryk University, Brno, Czech Republic.

The academic Web-as-Corpus

Adriano Ferraresi
DIT Department
University of Bologna
Forlì, Italy
adriano.ferraresi@unibo.it

Silvia Bernardini
DIT Department
University of Bologna
Forlì, Italy
silvia.bernardini@unibo.it

Abstract

As a result of the European Union's pressure towards internationalization, universities in many countries find themselves increasingly urged to provide information on their requirements and services and to promote themselves in English on the web. Hence the need for corpus resources and studies of institutional academic English used as an international language (or lingua franca) on the web. This paper introduces "acWaC-EU" (an acronym for "academic Web-as-Corpus in Europe"), a corpus of web pages in English crawled from the websites of European universities and annotated with contextual metadata. The corpus contains approximately 40 million words from native English universities and a similar number of words from universities based in all other European countries, in which English is used as a lingua franca. Thanks to the metadata, it is possible to regroup texts for comparison based, e.g., on the language family of the native language spoken in the country where the text was produced. The paper describes and evaluates the corpus construction pipeline and the corpus itself, presents a case study on the use of modal and semi-modal verbs in lingua franca vs. native texts, and looks at future developments, in particular as concerns simple heuristics for topic-/genre-oriented subcorpus construction.

1 Introduction

Attracting the best students, academics and researchers from outside the EU and favouring student and staff mobility across EU country borders are among the priorities of the European Higher Education Area, launched in 2010 as one of the achievements of the Bologna Process.[1] Loathed by some and enthusiastically endorsed by others, *internationalization* has become a buzzword of higher education across Europe.

Several studies investigating higher education policies have shown that for internationalisation to be successful, availability of academic modules and/or entire degree courses in English is essential (e.g. Altbach and Knight (2007)), and that one of the most effective means to reach a vast international audience is to publish (quality) contents in English on institutional websites, which are a primary source of information for up to 84% of prospective students (cf. Saichaie (2011, chapter 1) and references therein). As a result, one would expect a massive (and growing) presence on the web of institutional academic English from EU countries in which English is used as an international language. Considerable variability, however, is observed when taking into account the degree to which academic institutions of different European countries offer (web-based) English contents. In a large-scale study on Internet multilingualism, Callahan and Herring (2012) find that the presence of English as a secondary language is most widespread on the websites of West-European (and especially Scandinavian) universities, followed by universities from the post-Soviet bloc. On the other hand, Romance-language countries like France, Italy and Spain lag behind. Against this background, interventions aimed at supporting multilingualism in the institutional/administrative domain are therefore in order. On the practical/applied side, these may include the implementation of tools for assisting non-native writers in producing appropriate texts in this specialized domain (cf. Depraetere et al. (2011)); on the descriptive side, studies are required which shed light on the different commu-

[1] http://www.ehea.info/.

nicative strategies adopted by universities based in countries where English is used as a lingua franca (ELF) vs. a native language.

This paper aims to take steps towards filling this gap by introducing "acWaC-EU" (an acronym for "academic Web-as-Corpus in Europe"), a corpus of nearly 90-million words of web pages in English crawled from the websites of European universities. Section 2 reviews previous work, focusing in particular on the macro-genre sampled in the corpus, which we refer to as institutional academic language. Section 3 introduces the pipeline that was followed to build the corpus, describes an experiment that was devised to evaluate its efficacy and an evaluation of the final corpus make-up in terms of genres included in its ELF and native components. Section 4 presents a case study in which (semi-)modal verbs are compared across native and ELF subcorpora. Section 5 presents plans to increase usability of the corpus through the provision of a naive text classification relying on URL syntax, and Section 6 concludes by summing up the main issues covered in the paper and discussing future research directions.

2 Previous studies

By institutional academic language we refer to the wide range of texts used for everyday communication between higher education institutions and their stakeholders, which are likely to feature prominently on university websites – i.e. syllabi, course packs, welcome messages, mission statements, announcements, but also blogs, endorsements, press releases and so forth. Probably due to their subservient "housekeeping" function, these genres have so far been largely neglected as objects of study compared to the more central disciplinary genres (e.g. Ph.D. theses and defences, research articles (see Swales (2004) for an overview), and more recently book reviews (Römer, 2010), grant proposals (Connor and Upton, 2004), thesis acknowledgements, doctoral prize applications and bio statements (Hyland, 2011).

A few landmark works have however been produced mainly within applied (corpus) linguistics and critical discourse analysis. The former are motivated by the observation that institutional texts constitute a substantial share of the (non-research) texts that faculty members are expected to produce as part of their commitments (Hyon and Chen, 2004), as well as being required readings for students who need to "navigate the maze of university requirements and services" (Biber, 2006, 26). Biber (2006) provides a full-fledged account of the TOEFL 2000 Spoken and Written Academic Language corpus (T2K-SWAL), which includes both academic and institutional genres (e.g. handbooks, catalogues, programme web pages, course syllabi). The relevance of institutional genres for applied linguistics purposes is also endorsed by the builders of the MICASE corpus which includes, alongside more formal spoken academic registers, everyday events such as service encounters and campus tours (Simpson-Vlach and Leicher, 2006). Recent work has also begun to explore the specific features of different genres within the domain, e.g. course syllabi (Afros and Schryer, 2009; Gesuato, 2011) and the "About us" pages of university websites (Caiazzo, 2011).

Within the critical discourse analysis literature, work on the discursive practices of tertiary education institutions dates back to the seminal paper by Fairclough (1993, 143). Here, it was suggested that universities are "in the process of being transformed through the increasing salience within higher education of promotion as a communicative function". Surveying more recent trends in academic communication, Swales (2004, 9) argues that the "marketization" of university discourse has also been accompanied "by a shift in curricular perspective to the needs of the students (now seen as "customers") as opposed to the scholarly expectations of a discipline or the traditional offerings of a department". Evidence that this process is increasingly pervasive has been provided in a number of papers within critical discourse analysis. Mautner (2005, 38), for instance, shows how universities borrow commercial models, using persuasive style and "[l]exical imports from the business domain", a finding confirmed by Morrish and Sauntson (2013, 78), who argue that institutions "have adopted the language of business and industry, managerialism and neoliberalism".

Both the corpus linguistics and the critical discourse analysis works discussed so far have focused specifically on English texts produced by universities based in Anglophone countries. However academia is "one of those influential domains that have widely adopted English as their common language, and [...] where international communication characterizes the domain across the

board" (Mauranen, 2010, 21). Explorations of non-native English varieties in international academic settings have come under the focus of attention of scholars interested in ELF (see e.g. Jenkins (2011) for an overview). These studies have brought to the fore the importance of isolating the features that characterize effective communication in ELF, and set it apart from its native counterpart. To the best of our knowledge, the only study that has set out to compare ELF and native production in the institutional academic domain is Bernardini et al. (2010). The authors describe a corpus of institutional academic texts collected from the websites of British/Irish and Italian universities using a semi-automatic procedure that consisted in manually selecting relevant URLs and then using these as seeds for retrieving and downloading pages through the BootCaT toolkit (Baroni and Bernardini, 2004). The native and ELF subcorpora are compared in terms of genres and topics covered, phraseological patterns and stance expressions. Findings indicate that ELF texts are focused on spelling out instructions and requirements, while native texts promote institutions as service providers through a personal style.

Web-as-corpus works focusing on university websites include Rehm (2002), who built a corpus of German academic websites from which he extracted and analysed personal pages of academics, using the corpus as a testbed for developing an automatic genre classification method. Thelwall has conducted substantial work on university websites combining methods from web analytics and corpus linguistics. Apart from the methodological work reported on in Thelwall (2005a), work relevant to the present paper includes Thelwall (2005b), in which some basic textual features of university websites from Australia, New Zealand and the U.K. were contrasted. These included the relative number of known and unknown word types and the presence of high-frequency anomalies (i.e. frequent words not found in the BNC). Findings were used to draw conclusions about methods for clustering academic web documents. The language issue is also touched upon, though from a different perspective from ours. Thelwall (2005b) claims that "[a]n analysis of the university web sites of any mainland European country would need to separate out the pages written in different languages in order to get useful results. [...] Future scientific web intelligence research will need to take language factors into account".

3 Corpus construction and evaluation

3.1 The pipeline

The aim in building acWaC-EU was to obtain a large monolingual comparable corpus – a corpus setup widely used, e.g. in translation studies and studies of learner language – affording comparison of native and ELF varieties of English in the institutional academic domain, as attested on the websites of European universities. The pipeline that was developed to this end largely relies on off-the-shelf tools for cleaning and annotating web pages, but it implements a pre-crawling step addressing the non-trivial problem of automatically retrieving English contents in websites where English is not expected to be the main language. Unlike previous attempts in the web-as-corpus literature to retrieve contents in a specific language from multilingual websites (e.g. Resnik and Smith (2003) and Brunello (2012)), the pipeline avoids search engines, which ensures replicability of the corpus construction procedure.

The pipeline consists of three main steps: a) seed URL retrieval, b) harvesting of pages and c) post-hoc cleaning, annotation and indexing. In the first phase, a list of the URLs of all European universities is obtained from the Webometrics website, which publishes a yearly ranking of universities and other higher education institutions according to their presence on the web.[2] A Perl web crawler then visits each homepage and downloads it.

For universities based in countries where English is a native/official language (native universities for short), the URLs obtained from Webometrics are used to seed a second crawl. For all other European universities (ELF universities), the script analyses `<a>` tags within the HTML code looking for a link to an English-language (home)page. This is done by means of Regular Expressions matching the pattern `(english|eng|en)` (both lower- and uppercase) in the `href`, `class` and `title` attributes, and in link text.[3] The idea of exploiting link struc-

[2] `http://www.webometrics.info/`. The ranking also includes institutions which might not be considered as "proper" universities (e.g. independent research centres, music schools, etc.). No attempt was made to filter these out, as any decision as to what constitutes a "proper" university would have been highly arbitrary.

[3] The complete set of RegEx used can be found in the

ture to identify English pages also lies at the heart of the well-known STRAND algorithm proposed by Resnik and Smith (2003): its performance, however, heavily depended on the search engine used (i.e. Altavista), which only made it possible to take into account `href` attributes via the `inanchor` operator.

If no link is found through analysis of `<a>` tags, a second heuristic is used: if the `lang` or `content` attributes in the HTML header are set to `en`, `en-US` or `en-GB`, the page is signaled as "potentially in English". After preliminary inspection, it was decided that these pages had to be checked manually to discard false positives. Out of a total of 5,505 ELF university websites ranked by Webometrics, 2,622 "supposed" English homepages were found using the first method, and 236 using the second one: a check of a random sample of 200 URLs from the first set and of the 236 URLs in the second set revealed that 168 and 62 pages respectively were actually in English (corresponding to a precision of 84% and 26.3%).

The URLs found using the two heuristics and the native English homepages obtained from Webometrics are used to seed the second crawl. In this further step, the pages linked from the seed URLs are fetched, with two levels of recursion, i.e. we download pages if they are at most 2 links away from the seed URL. Of course, one could move deeper into the site structure, e.g. to increase corpus size, but this would be done at the expense of crawl efficiency: as shown in Section 3.2, as one moves away from the English homepage, contents in English dwindle. A total of 1,233,690 pages were downloaded (84% from ELF universities and 16% from native ones).

In the final phase, the crawled pages are cleaned using the tools developed for the web-derived, general-purpose WaCky corpora (Baroni et al., 2009), and in particular the language identifier, the boilperplate-stripping and de-duplication algorithms. After cleaning, the remaining pages are Part-of-Speech tagged using the TreeTagger and indexed for consultation with the Corpus Workbench. During this phase, contextual metadata are recorded with each text, including:

- URL of the web page and level in the site structure at which it was found (from 0 to 2, where 0 indicates the homepage);
- variety of English (native/ELF);
- name of the university which published the page and its rank according to the Webometrics classification;
- country where the university is based, European Union membership (yes/no), and language family of the official language spoken (e.g. France/Romance, Norway/Germanic, Russia/Slavic, etc.).

These metadata can be exploited for the construction of subcorpora, and form the basis for the analyses presented in Sections 3.3 and 4.

	ELF	**Native**	**TOTAL**
Tokens	41,696,310	46,172,429	87,868,739
Texts	73,296	68,011	14,1307
Universities	2,159	341	2,500
Countries	46	4	50

Table 1: acWaC-EU corpus statistics, by subcorpus.

Table 1 provides statistics about the corpus in its final, cleaned version. Notice that the size of the ELF and native subcorpora is roughly equivalent, both in terms of number of tokens and texts, even though the number of universities sampled is, as one would expect, much larger in the former than in the latter.

Additional information on acWaC-EU, e.g. the list of universities and countries sampled and the scripts that were used for its construction, are available from the page: `http://mrscoulter.sslmit.unibo.it/acwac/`. Work is under way to also make the corpus available through the same page (cf. Section 6).

3.2 Evaluating the pipeline

To assess its performance, the pipeline used to build acWaC-EU is compared to a baseline method whereby ELF university websites are crawled starting from their "initial" homepage in the respective local/national language. Other methods involving the use of search engines (e.g. the STRAND pipeline) were not taken into account, as they were not considered as viable alternatives to build acWaC-EU, for the reasons given in Section 3.1.

The methods are tested on a sample of 33 universities chosen randomly among those listed by

script available from `http://mrscoulter.sslmit.unibo.it/acwac/`.

Webometrics for 3 countries (for a total of 99 universities), i.e. Serbia, Spain and Sweden, whose official languages belong to one of the three language groups most represented in acWaC-EU, i.e. the Slavic, Romance and Germanic groups.

The acWaC-EU and baseline method differ only in terms of the page which is used to seed the crawl, i.e. the English homepage identified through step a) (cf. Section 3.1) vs. the homepage listed by Webometrics. For the purposes of this comparison, the crawl is performed with three levels of recursion, so as to compensate for any advantage deriving from the fact that the acWaC-EU pipeline starts from one level deeper into the site compared to the baseline. After crawling, pages are post-processed with the same tools used for acWaC-EU.

	Level0	Level1	Level2	Level3
ACWAC-EU METHOD				
Downloaded	73	3,771	42,070	275,638
Final	22	937	5,818	12,318
RATIO	30.1%	24.8%	13.8%	4.4%
BASELINE METHOD				
Downloaded	99	6,470	70,605	486,900
Final	0	133	2,396	12,767
RATIO	0.0%	2.1%	3.4%	2.6%

Table 2: Comparison of the acWaC-EU and baseline method.

Table 2 displays statistics about the performance of the two methods, measured in terms of the ratio of web pages preserved after language filtering and de-duplication out of the total number of pages downloaded at each level of crawling. Results indicate that the acWaC-EU pipeline achieves better performance at all levels of crawling, although after level 2 the proportion of pages preserved in the final corpus drops to a much smaller percentage, that gets close to the one obtained with the baseline method.

The two methods yield similar numbers of pages for the three countries sampled, i.e. between 76% and 79% for Sweden, between 17% and 20% for Spain and between 3% and 4% for Serbia. The number of universities contributing at least one page is slightly higher in the baseline corpus (83 vs. 70 out of 99 universities), but becomes roughly equivalent when only universities contributing more than 10 pages are considered (63 vs. 56).

Figure 1: Distribution of genres by subcorpus

3.3 Assessing corpus composition

As with any web-as-corpus pipeline, control over the final corpus composition is limited, and post-hoc checks are needed to ensure that corpus contents match the targeted population. In the case of acWaC-EU, this also means that its main components, i.e. the ELF and native subcorpora, should be roughly comparable.

To assess composition of acWaC-EU, a random sample of 99 documents was extracted from each of the native and ELF subcorpora; the latter sample was obtained by extracting 33 pages from each of the ELF-Germanic, ELF-Romance and ELF-Slavic subcorpora. The two authors read through all the documents and classified them independently in terms of (broad) topic/genre classes, revising jointly the cases where disagreement was observed. The categories were based on a previous effort to categorize English texts in the institutional academic domain (reported on in Bernardini et al. (2010)).

The results of the analysis are shown in Figure 1. The distribution of texts among the different categories is not identical in the four components of acWaC-EU under consideration: compared to the ELF subcorpora, the native one contains more descriptions of degree courses and modules ($\Rightarrow$ *deg*) and of facilities offered by schools and departments, including descriptions of web-based services, such as library catalogues ($\Rightarrow$ *fac*); it also features more pages related to students'

life, e.g. texts by and for current and prospective students on jobs, food and social life (⇒ *stu*). The distribution of text types which most resembles that of the native subcorpus is found in the ELF-Germanic one, which, however, contains more general texts describing the university/single faculties, and/or welcoming prospective students (⇒ *gen*). The ELF-Romance and ELF-Slavic subcorpora display the most dissimilar distribution both compared to the native and ELF-Germanic components and to each other: the ELF-Romance component features a larger proportion of research-related pages by individual academics and research teams (⇒ *res*), and of regulatory texts (e.g. entry requirements, offers of fellowships, etc.; ⇒ *reg*), a trend which was also observed by Bernardini et al. (2010) in their comparison of UK/Irish vs. Italian university websites; on the other hand, the ELF-Slavic component features the largest proportion of general texts and of announcements of academic news and events (⇒ *new*). The proportion of texts which could not be assigned to any other category and of web navigation pages (e.g. sitemaps and error messages; ⇒ *web/oth*), where nearly no university-related content is present, is consistently below 10%.

The picture that emerges is one where ELF and native subcorpora vary in terms of the relative proportions of text types, with native texts being more focused on aspects which are directly relevant to students (courses, facilities, student life), and ELF texts preferring to advertise themselves through general texts or by providing information on their research and/or academic events. The degree of variation observed, which is probably due to different communicative strategies and/or institutional backgrounds in the different countries, does not seem to hinder subcorpus comparability.

4 Case study: (Semi-)modal verbs

Modal (and semi-modal) verbs seem an obvious starting point in the comparison of communication strategies used by European Universities in their English-language websites, since they constitute "by far the most common grammatical device use to mark stance in university registers" (Biber, 2006, 95). Taken together, modals (as identified by the TreeTagger) are used more frequently in the native than in the ELF subcorpus (11,837 vs. 7,892 pmw). The same is true of the semi-modals *have to*, *be going to* and *need to* (747 vs. 620 pmw).

	ELF	**NAT**	***p***
can	2119.16	2611.95	<0.001
could	198.15	286.19	*ns*
have to	322.16	214.33	<0.001
may	613.82	1107.46	<0.001
might	89.60	147.19	*ns*
must	611.35	529.36	<0.001
need to	241.96	501.64	<0.001
shall	128.26	85.03	<0.001
should	579.12	765.89	<0.05
will	3193.54	5718.24	<0.001
would	347.75	571.25	*ns*

Table 3: Frequency pmw of (semi-)modal verbs in the ELF and native subcorpora, and *p*-values of difference (Fisher's exact test).

Looking more closely, the majority of the (semi-)modals tested (those with frequencies above 50 pmw in at least one subcorpus) are used significantly more frequently in the native subcorpus. This is the case with *will*, *can*, *may*, *should* and *need to*. *Must*, *shall* and *have to*, on the other hand, display significantly higher frequencies in the ELF subcorpus (cf. Table 3).[4]

The picture emerging from this comparison is one in which native texts seem to use modals of permission/possibility/ability (*can* and *may*) much more extensively than ELF texts. They also use *will* very often, a modal expressing prediction and volition. To express obligation and necessity, native texts favour the more indirect options offered by the language, namely *should* and *need to*. ELF texts instead have recourse to modals and semi-modals especially to express obligation/necessity, and when they do they favour the more direct forms (i.e. *must* and *have to*). The frequent use of *shall* is also noteworthy since, differently from *will* and coherently with the general picture, "it marks volition more often than prediction" (Biber et al., 1999, 495).

One could speculate that the differences observed do not apply to ELF vs. native texts, but might be due to other variables. The corpus metadata, allowing on the fly subcorpus construction according to different parameters, can be used to investigate this hypothesis. In Figure 2, we compare the normalized frequencies of each of the (semi-)modals analysed above in the native subcorpus and in three subcorpora of texts from the

[4] *Need*, *ought*, *had better*, *be supposed to* and *have got to* were not tested because they do not reach the threshold of 50 occurrences pmw.

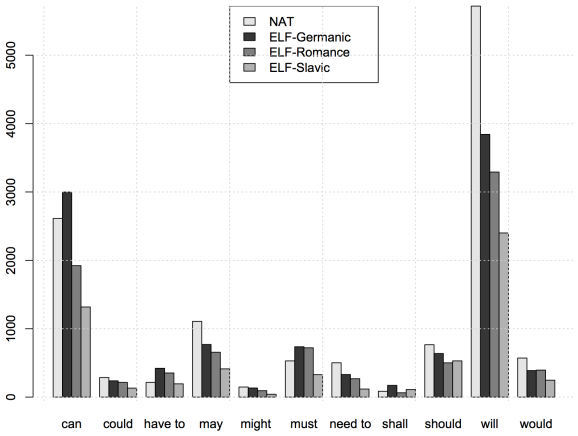


Figure 2: Distribution of (semi-)modal verbs by subcorpus

different language families (ELF-Germanic, ELF-Romance and ELF-Slavic).

The most common pattern is one in which the native texts contain more (semi-)modals than the Germanic texts, which in turn contain more (semi-)modals than Romance texts, which in turn contain more (semi-)modals than Slavic texts. With small local variations, this is true of *could*, *may*, *might*, *need to*, *should*, *will* and *would*. The preference for the direct expression of obligation/necessity (*must* and *have to*) is due to the larger use made of these verbs in Romance and Germanic ELF countries, whereas *can* and *shall* are favoured by universities based in ELF countries where Germanic languages are spoken. Slavic texts consistently use fewer (semi-)modals, with the exception of *shall*. While the higher frequency of *shall* in ELF texts was also observable in the general (ELF vs. native) comparison, and the language family comparison only served to highlight more fine-grained differences, in the case of *can* the initial observation does not hold: Germanic ELF texts use this modal *more* often than native texts, but since other language families use it substantially *less* often, the two figures level themselves out. In other words, differences in the use of *can* seem to be interpretable mainly in terms of first language background, whereas e.g. differences in the use of *will* (and most other modals) emerge as potential features of institutional academic ELF.

Although a more thorough analysis is beyond the scope of this paper, it is worth looking more closely at the co-text of *shall* in the different language family subcorpora, to highlight usage differences emerging alongside quantitative ones. A search for a personal pronoun or noun, followed by *shall*, followed by a verb lemma, returns the top 20 trigrams shown in Table 4. In native texts *shall* is used in formal, regulatory texts (cf. "personal data shall be processed", "the appeal shall be heard") or as an alternative for *will* with first person subjects, to mark personal volition ("we shall be offering", "we shall discuss topics"). Romance texts seem rather similar (cf. the presence of formal, regulatory terms like "litigation" and of first person subjects among the top collocates), whereas Germanic and to some extent Slavic texts seem to differ. Though formal, with extensive use of the passive, the co-texts do not hint at a regulatory function (cf. "At least one of the supervisors shall be employed", "employees shall be asked to evaluate"), and the first person pronouns are virtually absent from the top of the list of subject collocates. These qualitative differences between the use of *shall* in ELF-Germanic and native texts could only be highlighted through a closer examination of its usage patterns. The obvious next step would be to compare usage of *shall* in the different Germanic language countries. The corpus metadata make this type of investigation straightforward.

5 Future perspectives: genre- and topic-restricted subcorpora

In this paper we have considered university websites as wholes, seeing them as the means through which universities present/promote themselves and interact with their various stakeholders. Ideally, however, it would also be useful to compare (loosely-defined) genre- or topic-restricted subsets of native and ELF texts.

As a first attempt to define subsets of texts on the basis of external criteria, we used a simple heuristic based on URL syntax: a frequency list was generated for slash-separated parts of URLs, after removing transfer protocols and domain names. The top 3 items in the list, with respective frequencies, are "news" (8488), "courses" (5903) and "research" (5170). The sizes in tokens of the genre-based ELF and native subcorpora dealing with "news", "courses" and "research" is as shown in Table 5.

To check if these expressions are in fact effective cues as to the actual contents of the respective pages, 50 randomly selected URLs contain-

Native	ELF-Germanic	ELF-Romance	ELF-Slavic
student shall be	student shall be	agreement shall continue	student shall be
datum shall be	student shall demonstrate	form shall be	contest shall be
fee shall be	thesis shall be	it shall be	worker shall be
it shall be	student shall have	agreement shall be	staff shall be
we shall be	candidate shall be	datum shall be	study shall be
term shall be	it shall be	document shall be	it shall be
condition shall be	supervisor shall be	application shall be	document shall be
candidate shall be	examination shall be	we shall prepare	decision shall be
application shall be	student shall develop	we shall be	member shall be
member shall be	dissertation shall be	registration shall be	program shall be
I shall be	report shall be	student shall submit	it shall comply
you shall be	application shall be	student shall be	fee shall be
we shall discuss	student shall acquire	enrolment shall be	application shall be
we shall have	employee shall be	fee shall be	education shall be
appeal shall be	they shall be	you shall live	paper shall be
person shall be	activity shall be	it shall have	applicant shall have
it shall have	education shall involve	student shall have	we shall watch
provision shall be	supervisor shall have	they shall be	they shall be
matter shall be	grade shall be	tranche shall be	procedure shall be
we shall do	applicant shall have	litigation shall come	programme shall be

Table 4: 20 most frequent lemma sequences including the modal *shall*.

	research	courses	news
ELF	1,901,098	800,487	3,673,205
NAT	2,566,095	7,576,082	5,488,887

Table 5: Size information (tokens) of the genre-restricted subcorpora.

ing the words *research*, *courses* and *news* were extracted from the ELF subcorpus and the same number were extracted from the native subcorpus, for a total of 300 URLs. Manual browsing of the corresponding pages showed that, with minimal variation between the two subcorpora:

- 90% of URLs including the word *courses* returned pages describing different types of courses (academic modules, summer schools, degree courses), and the remaining 10% referred to course-related regulations or facilities.
- 100% of URLs including the word *news* contained different types of news (e.g. shorter or longer pieces about academic events, partnerships, new discoveries etc.).
- 99% of the URLs including the word *research* – the exception being a page not in English – referred to research-related topics such as groups, findings, projects, and grants; infrastructure and support; staff profiles; homepages of institutes. While consistent in subject domain, these pages are rather varied in terms of genre, but this applies equally to native and ELF texts.

Other items in the frequency list appearing in 1,000 URLs or more are *international*, *staff*, *alumni*, *departments*, *services*, *about* and *home*. On the basis of the promising results obtained for the top three items, we hypothesize that it could be possible to define thematically coherent and reasonably sized subcorpora using URL-derived wordlists. Comparisons of ELF and native language could thus also be conducted in more controlled settings than in the complete acWaC-EU.

6 Conclusions

In this paper we have introduced the acWaC-EU corpus and presented a case study comparing the frequency of modal and semi-modal verbs in native English and ELF language university websites in Europe. We have observed that, overall, modals and semi-modals are used more frequently in native than in ELF texts, and that the latter express obligation/necessity more directly than the former. We have also suggested that ELF is not a monolithic entity: universities from specific language families may have their own preferences, cf. *shall* in university websites from Germanic language countries.

Plans for the near future include testing the usefulness of our simple heuristic for defining topic- and genre-restricted subcorpora on the basis of URL syntax, and experimenting with more advanced techniques for text/genre classification. We also plan to make the corpus available in ways that do not infringe copyright laws, e.g. distributing it as a set of n-grams, along the lines of the Google Books n-gram dataset (Michel et al., 2011). Finally, the pipeline used to build acWaC-EU could be easily adapted to perform a new crawl on university websites worldwide.

Acknowledgments

We are grateful to the three anonymous reviewers of our extended abstract for their suggestions and to Federico Gaspari for reading and commenting on the full paper.

References

Elena Afros and Catherine F. Schryer. 2009. The genre of syllabus in higher education. *Journal of English for Academic Purposes*, 8(3):224–233.

Philip G. Altbach and Jane Knight. 2007. The internationalization of higher education: Motivations and realities. *Journal of Studies in International Education*, 11(3-4):290–305.

Marco Baroni and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*, pages 1313–1316, Lisbon, Portugal. ELDA.

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

Silvia Bernardini, Adriano Ferraresi, and Federico Gaspari. 2010. Institutional academic English in the European context: A web-as-corpus approach to comparing native and non-native language. In Ángeles Linde López and Rosalía Crespo Jiménez, editors, *Professional English in the European context: The EHEA challenge*, pages 27–53. Peter Lang, Bern.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Longman, London.

Douglas Biber. 2006. *University language: A corpus-based study of spoken and written registers*. John Benjamins, Amsterdam.

Marco Brunello. 2012. Understanding the composition of parallel corpora from the web. In Adam Kilgarriff and Serge Sharoff, editors, *Proceedings of the Seventh Web as Corpus Workshop (WAC7)*, pages 8–14.

Luisa Caiazzo. 2011. Hybridization in institutional language: Exploring *we* in the 'About us' page of university websites. In Srikant Sarangi, Vanda Polese, and Giuditta Caliendo, editors, *Genre(s) on the Move: Hybridization and Discourse Change in Specialized Communication*, pages 243–260. Edizioni Scientifiche Italiane, Napoli.

Ewa Callahan and Susan C. Herring. 2012. Language choice on university websites: Longitudinal trends. *International Journal of Communication*, 6:322–355.

Ulla Connor and Thomas A. Upton. 2004. The genre of grant proposals: A corpus linguistic analysis. In Ulla Connor and Thomas A. Upton, editors, *Discourse in the professions: Perspectives from corpus linguistics*, pages 235–256. John Benjamins, Amsterdam and Philadelphia.

Heidi Depraetere, Joachim Van den Bogaert, and Joeri Van de Walle. 2011. Bologna translation service: Online translation of course syllabi and study programmes in English. In Mikel L. Forcada, Heidi Depraetere, and Vincent Vandeghinste, editors, *Proceedings of the 15th conference of the European Association for Machine Translation*, pages 29–34, Leuven, Belgium, May.

Norman Fairclough. 1993. Critical discourse analysis and the marketization of public discourse: The universities. *Discourse & Society*, 4(2):133–168.

Sara Gesuato. 2011. Course descriptions: Communicative practices of an institutional genre. In Srikant Sarangi, Vanda Polese, and Giuditta Caliendo, editors, *Genre(s) on the Move: Hybridization and Discourse Change in Specialized Communication*, pages 221–241. Edizioni Scientifiche Italiane, Napoli.

Ken Hyland. 2011. Projecting an academic identity in some reflective genres. *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos*, 21:9–30.

Sunny Hyon and Rong Chen. 2004. Beyond the research article: University faculty genres and EAP graduate preparation. *English for Specific Purposes*, 23(3):233–263.

Jennifer Jenkins. 2011. Accommodating (to) ELF in the international university. *Journal of Pragmatics*, 43(4):926–936.

Anna Mauranen. 2010. Features of English as a lingua franca in academia. *Helsinki English Studies*, 6:6–28.

Gerlinde Mautner. 2005. For-profit discourse in the nonprofit and public sectors. In Guido Erreygers and Geert Jacobs, editors, *Language, communication and the economy*, pages 25–44. John Benjamins, Amsterdam.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Liz Morrish and Helen Sauntson. 2013. "Business-facing motors for economic development": An appraisal analysis of visions and values in the marketised UK university. *Critical Discourse Studies*, 10(1):61–80.

Georg Rehm. 2002. Towards automatic web genre identification: A corpus-based approach in the domain of academia by example of the academic's personal homepage. In *Proceedings of the 35th Hawaii International Conference on System Sciences*, pages 1143 – 1152.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Ute Römer. 2010. Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction*, 3(1):95–119.

Kem Saichaie. 2011. *Representation on college and university websites: An approach using critical discourse analysis*. Ph.D. thesis, University of Iowa.

Rita C. Simpson-Vlach and Sheryl Leicher. 2006. *The MICASE Handbook: A resource for users of the Michigan Corpus of Academic Spoken English*. The University of Michigan Press, Ann Arbor.

John Malcolm Swales. 2004. *Research genres. Explorations and applications*. Cambridge University Press, Cambridge.

Mike Thelwall. 2005a. Creating and using web corpora. *International Journal of Corpus Linguistics*, 10(4):517–541.

Mike Thelwall. 2005b. Text characteristics of English language university web sites. *Journal of the American Society for Information Science and Technology*, 56(6):609–619.

A Compact but Linguistically Detailed Database for German Verb Subcategorisation relying on Dependency Parses from Web Corpora: Tool, Guidelines and Resource

Silke Scheible, Sabine Schulte im Walde, Marion Weller, Max Kisselew
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
{scheible,schulte,wellermn,kisselmx}@ims.uni-stuttgart.de

Abstract

This paper describes the *SubCat-Extractor* as a novel tool to obtain verb subcategorisation data from parsed German web corpora. The SubCat-Extractor is based on a set of detailed rules that go beyond what is directly accessible in the parses. The extracted subcategorisation database is represented in a compact but linguistically detailed and flexible format, comprising various aspects of verb information, complement information and sentence information, within a one-line-per-clause style. We describe the tool, the extraction rules and the obtained resource database, as well as actual and potential uses in computational linguistics.

1 Introduction

Within the area of (automatic) lexical acquisition, the definition of lexical verb information has been a major focus, because verbs play a central role for the structure and the meaning of sentences and discourse. On the one hand, this has led to a range of manually or semi-automatically developed lexical resources focusing on verb information, such as the Levin classes (Levin, 1993), VerbNet (Kipper Schuler, 2006), FrameNet[1] (Fillmore et al., 2003), and PropBank (Palmer et al., 2005). On the other hand, we find automatic approaches to the induction of verb subcategorisation information at the syntax-semantics interface for a large number of languages, including Briscoe and Carroll (1997) for English; Sarkar and Zeman (2000) for Czech; Schulte im Walde (2002) for German; and Messiant (2008) for French. This basic kind of verb knowledge has been shown to be useful in many NLP tasks such as information extraction (Surdeanu et al., 2003; Venturi et al., 2009), parsing (Carroll et al., 1998; Carroll and Fang, 2004) and word sense disambiguation (Kohomban and Lee, 2005; McCarthy et al., 2007).

Subcategorisation information is not directly accessible in most standard annotated corpora, and thus typically requires a complex approach to induce verb knowledge at the syntax-semantic interface, cf. Schulte im Walde (2009) for an overview of methodologies. Even more, with the advent of web corpora, empirical linguistic researchers aim to rely on large corpus resources but have to face data where not only deep tools but also standard tools such as tokenisers and taggers often fail.

We describe a novel tool to extract verb subcategorisation data from parsed German web corpora. While relying on a dependency parser, our extraction was based on a set of detailed guidelines to maximise the linguistic value of the subcategorisation information but nevertheless represent the data in a compact, flexible format. In the following, we outline our subcategorisation extractor and describe the format of the subcategorisation database, as well as actual and potential uses in computational linguistics.

2 Subcategorisation Extraction: Tool, Rules and Resource Database

This section provides an overview of the *SubCat-Extractor*, a new tool for extracting verb subcategorisation information. The goal of the SubCat-Extractor is to extract verbs with their complements from parsed German data following a special set of extraction rules devised for this purpose.

[1] Even though the FrameNet approach does not only include knowledge about verbal predicates, they play a major role in the actual lexicons.

Position	Word	Lemma	POS	Morphology	Head	Dependency Relation
1	Er	er	PPER	nom, sg, masc, 3	2	SB (subject)
2	fliegt	fliegen	VVFIN	sg, 3, pres, ind	0	–
3	am	an	APPRART	dat, sg, neut	2	MO (modifier)
4	Wochenende	Wochenende	NN	dat, sg, neut	3	NK (noun kernel element)
5	nach	nach	APPR		2	MO (modifier)
6	Berlin	Berlin	NE	dat, sg, neut	5	NK (noun kernel element)
7	.	–	$.		6	–

Table 1: Example input.

In this section, we describe the input format for the SubCat-Extractor (Section 2.1), the specificities of the extraction rules (Section 2.2), the output format (Section 2.3) and the induced subcategorisation database (Section 2.4) in some detail.

2.1 Input Format

The input format required by the SubCat-Extractor is parsed text produced by Bernd Bohnet's MATE dependency parser (Bohnet, 2010). The parses are defined according to the tab-separated CoNNL[2] format, so in principle any parser output in CoNNL format can be processed by the SubCat-Extractor. Since the extraction rules rely on part-of-speech and syntactic function information in the parses, the respective format specifications have to be taken into account, too: The SubCat-Extractor tool is specified for part-of-speech tags from the *STTS* tagset (Schiller et al., 1999) and syntactic functions from *TIGER* (Brants et al., 2004; Seeker and Kuhn, 2012).

Table 1 shows an example sentence from the Bohnet parser that can serve as input to the SubCat-Extractor: *Er fliegt am Wochenende nach Berlin.* 'He flies to Berlin at the weekend'. For simplicity, we omit columns that consistently do not carry information: in the actual parser output, some columns used for evaluation purposes do not provide information for our parsing purposes. Accordingly, the information in Table 1 is restricted to the following information: the first column shows the sentence position, the second column shows the actual word type, the third column shows the lemma, the forth column shows the part-of-speech, the fifth column shows morphological information, the sixth column shows the head of the dependency relation, and the seventh column specifies the dependency relation, i.e., the syntactic function. For applying the SubCat-Extractor, each sentence must be followed by an empty line.

[2] www.clips.ua.ac.be/conll/

2.2 Extraction Rules

The SubCat-Extractor considers any verbs that are POS-tagged as finite (V*FIN), infinite (V*INF), or participial (V*PP) in the input files. We have devised detailed rules to extract the subcategorisation information, going beyond what is directly accessible in the parses. In particular, our rules include the following cases:

- Identification of relevant dependants of finite full verbs, across tenses.
- Identification of the auxiliaries *sein* 'to be' and *haben* 'to have' and modal verbs as full verbs, excluding all other instances from consideration.
- Identification of relevant dependants of infinite verb forms occurring with finite auxiliaries/modals.
- Distinguishing between active/passive voice.
- Resolving particle verbs.

An example of only indirectly accessible information in the parses is the definition of subjects, which –in the parses– are always attached to the finite verb; so in a sentence like *Die Mutter würde Suppe machen.* 'The mother might make soup.' we have to induce that *Mutter* 'mother' is the subject of *machen* 'make' because it is not a dependant of the full verb.

Appendix A provides more details of our rules, which represent the core of the SubCat-Extractor, showing under which conditions the rules apply, and what information is extracted. The list might serve as guidelines for anyone interested in applying or extending the SubCat-Extractor. Examples of the rules can be found in Appendix B. The complete guidelines are available from www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/subcat-extractor.en.html.

2.3 Output Format

The output of the SubCat-Extractor represents a compact but linguistically detailed database for German verb subcategorisation: It contains the extracted verbs along with the following tab-separated information:

(1) verb information;
(2) subcategorisation information;
(3) applied rule;
(4) whole sentence.

In the following, this information is described in more detail.

(1) Verb Information: Information on the extracted target verb consists of the following four parts (separated by colons):

1. *Dependency relation of the target verb*, according to the TIGER annotation scheme. For verbs located at the root position of the parse the relation is '–'. If the verb is included in a passive construction, the relation is prefixed with the label 'PAS_'.
2. *Part-of-speech (POS) tag* of the target verb, according to STTS.
3. *Position* of the target verb in the sentence, with the count starting at zero.
4. *Lemma* of the target verb.

Examples of this verb information are

- `--:VVFIN:2:planen`
- `OC:VVPP:4:entscheiden`
- `PAS_OC:VVINF:9:beantworten`

Of special interest concerning German particle verbs is the following specification: In cases where the SubCat-Extractor locates a verb particle in the sentence (with dependency relation 'SVP') that directly depends on the target verb, the particle is added as prefix to the lemma. An example of this procedure is *Petacchi schied verletzt aus.* → `--:VVFIN:1:ausscheiden`.

(2) Subcategorisation Information: The subcategorisation contains all complements $\texttt{Comp}_i$ of a given target verb as determined by the extraction rules in Appendix A, disregarding the distinction between arguments and adjuncts. The information is listed within angle brackets, and individual complements are separated by pipe symbols:

$\texttt{<Comp}_1\texttt{|Comp}_2\texttt{|...|Comp}_n\texttt{>}$

The complements included in the subcategorisation information are distinguished as follows:

(a) ***All complements (but PPs):*** SB (subject), EP (expletive), SBP (passivised subject), MO (modifier; restricted to adverbs), OA (accusative object), OA2 (ditto, in case there are two OAs in the same clause), OC (clausal object), OG (genitive object), PG (phrasal genitive), DA (dative object), PD (predicate), NG (negation), and AG (genitive attribute) use the same format as the verb information described above, i.e.

1. Dependency relation of the complement.
2. POS tag of the complement.
3. Position of the complement in the sentence.
4. Lemma of the complement.

An example complement (a subject represented by a personal pronoun (PPER), *ich* 'I') would be `SB:PPER:8:ich`.

An important feature of the subcategorisation extraction is that any subject (SB) tagged as relative pronoun (PRELS) is resolved to its ancestor, for example: *Kinder, die müde sind, ...* ('children who are tired, ...') → `<SB:NN:0:Kind|PD:ADJD:3:müde>`.

(b) ***PPs:*** MO (modifier; excluding adverbs), MNR (postnominal modifier), and OP (prepositional object with POS tag APPR (preposition) or APPRART (preposition incorporating article)), as well as CVC (collocational verb construction) introduce prepositional phrases (PPs). For this reason, the individual entries are further extended by adding the arguments of the prepositions. Double colons are used to separate preposition information from PP argument information:

1. Dependency relation of the preposition.
2. POS tag of the preposition.
3. Position of the preposition in the sentence.
4. Lemma of the preposition.

double colon ::

5. POS tag of the PP argument.
6. Case of the PP argument.
7. Position of the PP argument.
8. Lemma of the PP argument.

An example PP complement (*im Sommer* 'in the summer') would be `MO:APPRART:6:in::NN:dat:7:Sommer`.

Verb Information	Subcategorisation Information & Sentence(s)
–:VVFIN:1:fliegen	<SB:PPER:0:er|MO:APPRART:2:an::NN:dat:3:Wochenende|MO:APPR:4:nach::NE:dat:5:Berlin> *[Er]SB [[fliegt]]– [am]MO Wochenende [nach]MO Berlin .*
–:VVFIN:1:ausscheiden	<SB:NE:0:Petacchi|MO:VVPP:2:verletzen> *[Petacchi]SB [[schied]]– [verletzt]MO:OTHER [[aus]]SVP .*
–:VVFIN:2:stattfinden	<SB:NN:1:Kulturfestival|MO:APPRART:3:in::NN:dat:4:Sommer> *Zahlreiche [Kulturfestivals]SB [[finden]]– [im]MO Sommer [[statt]]SVP .*
OC:VVINF:6:verstehen	<SB:PIS:1:man|OA:NN:3:Begriff|CP:KOUS:0:wenn> *Wenn man den [Begriff]OA der Netzwerkeffekte [[verstehen]]OC *will*, . . .*
OC:VVPP:6:fahren	<SB:NE:1:Zabel|MO:ADV:3:gerne|MO:APPR:4:in::NE:dat:5:Gelb> *Erik Zabel ∗wäre∗ [gerne]MO:ADV [in]MO Gelb [[gefahren]]OC [. . .]* *Erik [Zabel]SB [[wäre]]– gerne in Gelb gefahren [. . .]*
PAS_OC:VVPP:5:kaufen	<SB:NN:0:Tier|MO:APPR:2:aus::NN:dat:4:Grund> *Tiere *werden* [aus]MO verschiedensten Gründen [[gekauft]]OC .*

Table 2: Example output.

If a PP involves coordination, both parts are resolved and included. For example: *im Sommer und Winter* induces `MO:APPRART:6:in::NN:dat:7:Sommer|MO:APPRART:6:in::NN:dat:9:Winter`.

(c) ***Conjunctions:*** The conjunction POS tags KON, CJ, CD, and – are excluded from consideration. The PPs are an exception to this (see above).

(3) Applied Rule: The rule that was applied to extract the verb and subcategorisation information is denoted, cf. Appendix A.

(4) Sentence: Finally, the whole sentence in which the target verb occurs is listed with the following mark-up:

- Double brackets *[[. . .]]* denote the verb.
- Single brackets followed by a label *[. . .]LABEL* denote complements of the target verb and their dependency relations.
- Curly brackets {. . . } denote the parent of the target verb.
- Asterisks ∗. . . ∗ are used to mark up a finite (auxiliary) verb on which the target verb depends and whose complements are added to the target verb's subcategorisation.
- Whenever the dependants of a finite (auxiliary) verb are included in the frame, a second sentence is added to the output showing the dependants of the respective finite verb (see example *fahren* in Table 2).

Examples Table 2 provides examples of the subcategorisation output, including those mentioned in the preceding parts of this section.

2.4 Subcategorisation Resource

So far, we have applied the SubCat-Extractor to dependency parses of the German web corpus *sdeWaC* (Faaß and Eckart, 2013),[3] a cleaned version of the German web corpus *deWaC* created by the *WaCky* group (Baroni et al., 2009). The corpus cleaning had focused mainly on removing duplicates from the *deWaC*, and on disregarding sentences that were syntactically ill-formed (relying on a parsability index provided by a standard dependency parser (Schiehlen, 2003)). The sdeWaC contains approx. 880 million words and is provided by `wacky.sslmit.unibo.it/`.

The *sdeWaC* subcategorisation database comprises 73,745,759 lines (representing the number of extracted target verb clauses). 63,463,223 (86%) of the target verb tokens appeared in active voice, and 10,282,536 (14%) of them appeared in passive voice. Table 3 shows the distribution of the verb clauses over full, auxiliary and modal verbs.

POS	Number of Clauses
VAFIN	11,395,914
VAINF	901,106
VAPP	302,586
VMFIN	348,056
VMINF	4,373
VMPP	5,959
VVFIN	33,640,028
VVINF	11,410,381
VVIZU	1,129,094
VVPP	14,608,262

Table 3: Full, auxiliary and modal verb clauses.

[3] `www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/sdewac.en.html`

3 Applications

The subcategorisation extraction tool and –more specifically– the subcategorisation resource described in the previous section are of great potential use because the information is represented in a compact format, but nevertheless with sufficient details for many research questions. Furthermore, the linear one-line-per-clause format allows quick and easy access to the data; in many cases, basic unix tools or simple perl or python scripts can be used, rather than going into the complexity of parse structures for each research question. The following paragraphs introduce applications of the database within our research project.

Subcategorisation Frame Lexicon As a natural and immediately subsequent step, we induced a subcategorisation frame lexicon from the verb data. Taking voice into account, we summed over the various complement combinations a verb lemma appeared with. For example, among the most frequent subcategorisation frames for the verb *glauben* 'believe' are a subcategorised clause 'believe that' (freq: 52,710), a subcategorised prepositional phrase with preposition an_{acc} 'believe in' (freq: 4,596) and an indirect object 'trust s.o.' (freq: 2,514). In addition, we took the actual complement heads into account. For example, among the most frequent combinations of heads that are subjects and indirect objects of *glauben* are <*man, Umfrage*> 'one, survey' and <*keiner, ihm*> 'nobody, him'. Paying attention to a specific complement type (e.g., the direct object within a transitive frame), we induced information that is relevant for collocation analyses. For example, among the most frequent indirect objects of *glauben* in a transitive frame are *Wort* 'word', *Bericht* 'report', and *Aussage* 'statement'. The subcategorisation frame lexicon has not been evaluated by itself but by application to various research studies (see below).

Subcategorisation Information for Statistical Machine Translation (SMT) Weller et al. (2013) is an example of research that applied our subcategorisation data. They improved the prediction on the case of noun phrases within an SMT system by integrating quantitative information about verb subcategorisation frames and verb–complement syntactic strength.

Prediction of Passives-of-Reflexives Zarries et al. (2013) exploited the linguistic and formatting advantages of our data, when they predicted the potential of building 'passives of reflexives' for German transitive verbs, such as

Erst wird sich$_{REFL}$ geküsst, ...
'First is REFL kissed, ...'.

They used the one-line-per-clause format to identify relevant subcategorisation frames of verbs and to restrict the types of noun complement heads that were allowed for specific syntactic functions.

Classification of Prototypical vs. Metaphorical Uses of Perception Verbs David (2013) used the subcategorisation information and the sentence information for (i) a manual inspection, (ii) corpus-based annotation and (iii) an automatic classification of prototypical vs. metaphorical uses of a selection of German perception verbs. The sentence information (cf. Section 2.3) in connection with the compact verb and subcategorisation supported the annotation purposes of perception verb senses; the verb information and the subcategorisation information were exploited as classification features. Relying on our subcategorisation database, a Decision Tree classification resulted in 55-60% accuracy scores in the 3-way and 4-way classifications.

Potential Uses In order to illustrate the potential of the information provided by our subcategorisation database, we add ideas of potential uses.

- *Complement order variations with regard to the verb type, the clause type and the subcategorisation frame:*
  The one-line-per-clause format provides verb information regarding the verb dependency and the position of the verb in the clause, as well as types and positions of the various complements, so it should be straightforward to quantify over the complement order variations ('scrambling') in relation to the verb information.
- *Extraction of light-verb constructions ('Funktionsverbgefüge') with prepositional objects:*
  The eight-tuple information in combination with the verb information should enable an easy access to light-verb constructions, as all relevant information is within one line of the subcategorisation database.

- *Quantification of verb modalities:*
  Since the information of whether a full verb depends on a modal verb (or not) is kept in the sentence information, the subcategorisation database should be useful to explore and quantify the modal conditions of verb types (in combination with specific types of complement heads).

4 Discussion

Section 2 introduced the SubCat-Extractor as a new tool for extracting verb subcategorisation information. The goal of the SubCat-Extractor is to extract German verbs along with their complements from parsed German data in tab-separated CoNNL format. We have devised detailed rules to extract the subcategorisation information from the dependency relations, going beyond what is directly accessible. So far, we have applied the SubCat-Extractor to dependency parses of a German web corpus, sdeWaC, comprising approx. 880 million words. Section 3 provided some actual and potential uses of the subcategorisation data.

The SubCat-Extractor is, of course, not restricted to be used for parses of only corpora from the web. It can be applied to any kind of corpus data, given that the corpus data is parsed by a parser with CoNNL format output, using the STTS tagset and the TIGER node set. We however defined the rules of the SubCat-Extractor in such a way that they are robust towards a large amount of noise in the underlying data. Since the MATE parser would always generate a parse for a sentence, and integrate erroneous as well as correct words and phrases, the rules of the SubCat-Extractor need to ensure a reliability filter for erroneous dependencies. For example, the sdeWaC web corpus parses commonly identify more than one subject for a full verb, because complement inflections (and thus case prediction) might be erroneous. Our rule set aims to extract at most one subject per full verb. In sum, we presented

- a ***new tool*** *(SubCat-Extractor)* that can be applied to German dependency parses and should be robust to extract verb subcategorisation information from web corpora,
- a ***new verb subcategorisation database*** obtained from the sdeWaC, with compact but nevertheless linguistically detailed information, and
- a ***new subcategorisation frame lexicon*** induced from the subcategorisation database.

The tool, the subcategorisation database and the subcategorisation frame lexicon are freely available for education, research and other non-commercial purposes:

- *tool:*
  www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/subcat-extractor.en.html
- *database/lexicon:*
  www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/subcat-database.en.html

Acknowledgements

The research presented in this work was funded by the DFG Research Project *Distributional Approaches to Semantic Relatedness* (Max Kisselew, Silke Scheible, Marion Weller), and the DFG Heisenberg Fellowship SCHU-2580/1-1 (Sabine Schulte im Walde).

Appendix A. Extraction Rules.

The *SubCat-Extractor* rules specify (i) the types of verbs that are considered for extraction, and (ii) the dependants of these verbs that are included in the subcategorisation information.

1) Extraction of finite verbs

Extraction rules for the finite verb types VVFIN **(a)**, VMFIN **(b)**, and VAFIN **(c)**:

Conditions:

- **(a)**: No special conditions.
- **(b)**: VMFIN does not depend on a V* (i.e. VMFIN is a full verb).
- **(c)**: VAFIN does not depend on a V* (i.e. VAFIN is a full verb).
  Special case (c'): a PD (predicate) depends on VAFIN.

Extract:

- **(a)**, **(b)**, **(c)**: All dependants of the finite verb.
- **(c')**: Also extract all dependants of PD as complements of VAFIN.

2) EXTRACTION OF PARTICIPLE VERBS

For V*PP we distinguish between four cases:

i) Compound tense: VVPP (a), VMPP (b), and VAPP (c) are extracted if the following applies:

Conditions:

- **(a)**, **(b)**, **(c)**: The sentence contains a finite verb (VAFIN or VMFIN).
- **(a)**, **(b)**, **(c)**: The participle verb is not a PD.
- **(a)**, **(b)**, **(c)**: The participle verb directly depends on a VA* whose head is *sein* 'to be' or *haben* 'to have'.
- **(b)**, **(c)**: There is no V* in the sentence that depends on the VMPP/VAPP.

Extract:

- **(a)**, **(b)**, **(c)**: All dependants of the participle verb and all complements of the finite verb.

ii) Passive: VVPP (a), VMPP (b), and VAPP (c) are extracted if the following conditions apply, and the participle verb is marked as passive.

Conditions:

- **(a)**, **(b)**, **(c)**: The sentence contains a finite verb (VAFIN or VMFIN).
- **(a)**, **(b)**, **(c)**: The participle verb is not a PD.
- **(a)**, **(b)**, **(c)**: The participle verb directly depends on a VA* whose head is *werden*.

Extract:

- **(a)**, **(b)**, **(c)**: All dependants of the participle verb and all complements of the finite verb.

iii) Past participle dependent on full verb: VVPP is extracted if the following conditions apply, and the participle verb is marked as passive.

Conditions:

- The sentence contains a finite verb.
- The participle verb is not a PD.
- The participle verb directly or indirectly depends on the finite verb.
- The participle verb directly depends on a full verb VV*.

Extract:

- All dependants of the participle verb and all complements of the finite verb.

iv) Predicative pronoun: Predicative pronouns are extracted if the following conditions apply, and the participle verb is marked as passive.

Conditions:

- The sentence contains a finite verb.
- The participle verb is a PD.
- The participle verb directly or indirectly depends on the finite verb.

Extract:

- All dependants of the participle verb and all complements of the finite verb.

3) EXTRACTION OF INFINITIVAL VERBS

For V*INF we distinguish two cases:

i) V*INF without *zu*, in combination with a modal verb or in a compound tense (future): VVINF **(a)**, VMINF **(b)**, and VAINF **(c)** are extracted as follows:

Conditions:

- **(a)**, **(b)**, **(c)**: Sentence contains a finite verb.
- **(a)**, **(b)**, **(c)**: V*INF has no particle *zu*.
- **(a)**, **(b)**, **(c)**: V*INF directly depends on VM* or VA* with head *werden*.
- **(b)**, **(c)**: The sentence does not contain a V* that depends on VMINF or VAINF.

Extract:

- **(a)**, **(b)**, **(c)**: All dependants of V*INF.
- **(a)**, **(b)**, **(c)**: All complements of V*INF.

ii) V*INF with *zu*: VVINF **(a)**, VMINF **(b)**, and VAINF **(c)** are extracted as follows:

Conditions:

- **(a) (b)**, **(c)**: Sentence contains a finite verb.
- **(a) (b)**, **(c)**: V*INF has a particle *zu*.
- **(a) (b)**, **(c)**: V*INF directly depends on a VV* or a VA*.
  Special case (ii'): VA* has head *sein*.
- **(b)**, **(c)**: The sentence does not contain a verb V* that depends on VMINF/VAINF.

Extract:

- **(a)**: All dependants of V*INF.
- **Special case (ii'):** Complements of the finite verb.

In the case of ii', V*INF is marked as passive.

Appendix B. Rule Examples.

Rule	Category	Examples	Glosses
Finite verbs			
1 (a)	VVFIN	Er *fliegt* am Wochenende nach New York. Das Kind *singt* schon seit Stunden. Sie *kauften* sich drei Blumen.	He *flies* to New York at the weekend. The child has been *singing* for hours. They *bought* (themselves) three flowers.
1 (b)	VMFIN	Er *will* das Auto. Er *darf* das bestimmt nicht.	He *wants* the car. He *may* certainly not.
1 (c)	VAFIN	Das Kind *hat* viele Autos. Peter *ist* im Kindergarten.	The child *has* many cars. Peter *is* in the kindergarden.
1 (c')	VAFIN	Die Eltern *sind* am meisten *betroffen*. Gegen ihn *ist* Anklage *erhoben* wegen ... Sie *waren* so *geliebt*.	The parents *are affected* the most. He *is charged* with ... They *were* so *beloved*.
Participle verbs: compound tense			
2 (i)(a)	VVPP	Die Mutter hat die Suppe *gekocht*. Die Mutter muss die Suppe *gekocht* haben. Das Kind ist weit *geschwommen*. Das Kind wird weit *geschwommen* sein.	The mother has *cooked* the soup. The mother must have *cooked* the soup. The child has *swum* far. The child will have *swum* far.
2 (i)(b)	VMPP	Er hat das unbedingt *gewollt*.	He absolutely *wanted* this.
2 (i)(c)	VAPP	Das Kind wird viele Autos *gehabt* haben. Peter wird im Kindergarten *gewesen* sein.	The child will have *had* many cars. Peter will have *been* in the kindergarden.
Participle verbs: passive			
2 (ii)(a)	VVPP	Die Suppe wird *gekocht*. Die Suppe soll *gekocht* werden. Die Suppe hat *gekocht* werden müssen.	The soup is being *cooked*. The soup should be *cooked*. The soup has had to be *cooked*.
Participle verbs: past participle dependent on full verb			
2 (iii)(a)	VVPP	Wir fühlen uns davon *betroffen*. Die Sachen gehen immer *verloren*.	We feel *affected* by that. The things always *get lost*.
Participle verbs: predicative pronoun			
2 (iv)[4]	V*PP	Die Eltern sind am meisten *betroffen*. Die Eltern bleiben am meisten *betroffen*. Sie waren so *geliebt*.	The parents are *affected* the most. The parents remain *affected* the most. They were so *beloved*.
Infinitival verbs without particle *zu*			
3 (i)(a)	VVINF	Er will *gehen*. Er darf sich das Auto morgen *kaufen*.	He wants to *go*. He may *buy* (himself) the car tomorrow.
3 (i)(b)	VMINF	Er wird das morgen *dürfen*. Er will das morgen *dürfen*.	He will *be allowed* (to do) this tomorrow. He wants to *be allowed* (to do) this tomorrow.
3 (i)(c)	VAINF	Er darf das Auto morgen *haben*. Er will morgen rechtzeitig da *sein*.	He may *have* the car tomorrow. He wants to *be* there in time tomorrow.
Infinitival verbs with particle *zu*			
3 (ii)(a)	VVINF	Er entscheidet zu *gehen*. Er hat gestern entschieden zu *gehen*. Er hat ihm befohlen zu *gehen*.	He decides to *leave*. Yesterday, he decided to *leave*. He told him to *leave*.
3 (ii)(b)	VMINF	Er hat sich entschieden mehr Inhalte zu *wollen*.	He decided to *want* more content.
3 (ii)(c)	VAINF	Er hat sich vorgenommen Zeit zu *haben*. Er hat vorgeschlagen dabei zu *sein*.	He intended to *have* time. He suggested to *be* there.
3 (ii')	V*INF	Die Hinweise sind zu *beachten*. Die Frage ist leicht zu *beantworten*. Die Hilfsmittel sind da zu *sein*.	The indications are to be *respected*. The question is easy to *answer*. The tools are to *be* there.

Table 4: Examples of sentences and applied rules.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Bernd Bohnet. 2010. Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97, Beijing, China.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation*, 2(4):597–620.

Ted Briscoe and John Carroll. 1997. Automatic Extraction of Subcategorization from Corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 356–363, Washington, DC.

John Carroll and Alex C. Fang. 2004. The Automatic Acquisition of Verb Subcategorisations and their Impact on the Performance of an HPSG Parser. In *Proceedings of the 1st International Joint Conference on Natural Language Processing*, pages 107–114, Sanya City, China.

John Carroll, Guido Minnen, and Ted Briscoe. 1998. Can Subcategorisation Probabilities Help a Statistical Parser? In *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora*, pages 118–126, Montreal, Canada.

Benjamin David. 2013. Deutsche Wahrnehmungsverben: Bedeutungsverschiebungen und deren manuelle und automatische Klassifikation. Studienarbeit. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Gertrud Faaß and Kerstin Eckart. 2013. SdeWaC – a Corpus of Parsable Sentences from the Web. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*, Darmstadt, Germany. To appear.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16:235–250.

Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania, Computer and Information Science.

Upali S. Kohomban and Wee Sun Lee. 2005. Learning Semantic Classes for Word Sense Disambiguation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 34–41, Ann Arbor, MI.

Beth Levin. 1993. *English Verb Classes and Alternations*. The University of Chicago Press.

Diana McCarthy, Sriram Venkatapathy, and Aravind K. Joshi. 2007. Detecting Compositionality of Verb-Object Combinations using Selectional Preferences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 369–379.

Cédric Messiant. 2008. A Subcategorization Acquisition System for French Verbs. In *Proceedings of the Student Research Workshop at the 46th Annual Meeting of the Association for Computational Linguistics*, pages 55–60, Columbus, OH.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated Resource of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Anoop Sarkar and Daniel Zeman. 2000. Automatic Extraction of Subcategorization Frames for Czech. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 691–697, Saarbrücken, Germany.

Michael Schiehlen. 2003. A Cascaded Finite-State Parser for German. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 163–166, Budapest, Hungary.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen, 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, and Seminar für Sprachwissenschaft, Universität Tübingen.

Sabine Schulte im Walde. 2002. A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In *Proceedings of the 3rd Conference on Language Resources and Evaluation*, volume IV, pages 1351–1357, Las Palmas de Gran Canaria, Spain.

Sabine Schulte im Walde. 2009. The Induction of Verb Frames and Verb Classes from Corpora. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook.*, volume 2 of *Handbooks of Linguistics and Communication Science*, chapter 44, pages 952–971. Mouton de Gruyter, Berlin.

Wolfgang Seeker and Jonas Kuhn. 2012. Making Ellipses Explicit in Dependency Conversion for a German Treebank. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 3132–3139, Istanbul, Turkey.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using Predicate-Argument Structures for Information Extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 8–15, Sapporo, Japan.

Giulia Venturi, Simonetta Montemagni, Simone Marchi, Yutaka Sasaki, Paul Thompson, John McNaught, and Sophia Ananiadou. 2009. Bootstrapping a Verb Lexicon for Biomedical Information Extraction. In Alexander Gelbukh, editor, *Linguistics and Intelligent Text Processing*, pages 137–148. Springer, Heidelberg.

Marion Weller, Alex Fraser, and Sabine Schulte im Walde. 2013. Using Subcategorization Knowledge to improve Case Prediction for Translation to German. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria. To appear.

Sina Zarrieß, Florian Schäfer, and Sabine Schulte im Walde. 2013. Passives of Reflexives: A Corpus Study. Talk at the International Conference *Linguistic Evidence 2013 – Berlin Special: Empirical, Theoretical and Computational Perspectives*, Humboldt-Universität zu Berlin, Germany.

Thug Breaks Man's Jaw: A Corpus Analysis of Responses to Interpersonal Street Violence

Andrew Brindle
St. John's University,
Department of Applied English,
499 Tam King Road, Tamsui, New Taipei City, Taiwan
andrewbr@mail.sju.edu.tw

Abstract

This paper examines a corpus of online responses to an article in an online edition of the British tabloid newspaper *The Sun* describing an act of interpersonal street violence between two men. The article produced 190 responses from readers, which were collected and compiled into a corpus that contained 6,606 tokens. Employing a corpus-driven approach, the data was investigated by undertaking concordance analyses of keywords and collocates of those words. The data was further analysed by taking into consideration multimodal information such as user names and avatar images in order to examine the significance of stating gender in correlation with views expressed. The findings indicate that regardless of the negative depiction of the aggressor in the article, the assailant and his actions were defended by certain posters, and at times admired and praised, while the victim was criticised for his lack of fighting skills, and not considered as innocent. However, the data also revealed that other respondents rejected such violence, demonstrating a continuum of reactions among the tabloid readership who responded to the article. The study found a marked difference of stances between those who stated that they were male to those who did not. The paper concludes by discussing the hypothesis that masculine identity and specifically hegemonic masculinity is constructed from multiple identities. Furthermore, the importance of investigating and analysing online peer groups is emphasised as an invaluable source in comprehending aspects of social behaviour within contemporary society.

Keywords: masculinities; interpersonal violence; corpus; discourse; online peer groups.

1. Introduction

The Internet and other Web-derived data have become a vast resource for corpus linguistics and natural language processing. In this study, texts of computer-mediated communication (CMC) taken from a message board of an online edition of a British tabloid newspaper, *The Sun,* were built into a corpus and analysed. The study researches the responses readers posted to an article in the newspaper which detailed an act of interpersonal street violence between two men in which one man was seriously injured. This research utilises a corpus-driven approach in order to discuss the attitudes articulated by the posters towards the act of violence, which it is argued, reflects upon the virtual identity of the posters.

As a result of the anonymity and freedoms of time and space, virtual identity is thought of as more unstable, performed and fluid than 'real' identity (Benwell and Stokoe, 2006), yet such a definition has similar qualities to postmodern identity which is described as both constructed and discursive (Bauman, 2007). Thus, this analysis aims to not only highlight the posters' attitudes towards such violence, but furthermore, demonstrate traits of identity through discursive accomplishment.

The theme of the discussion board is centred on an act of violence between two men. A great deal of what is bad in the world, from genocide to interpersonal violence, is the product of men and their masculinities (DeKeseredy and Schwartz, 2005). Work by criminologists such as Anderson (1990) have argued that instances of interpersonal violence originate from strongly held values in the construction and defence of personal street status

and that violence is a tool for both the formation of and the protection of self-image. Furthermore, Messerschmidt (2004) writes that among certain men violence is a core component of masculinity and a means of proving one's manhood. However, Winlow (2001) considers that street and pub fights function as a means for working-class men to actualise a masculine identity due to the loss of traditional industrial job opportunities in a postmodern society. Clearly, violence is one means by which certain men live up to the ideals of hegemonic masculinity; such practices may be learned through interactions with particular peer groups, or virtual peer groups, and understood as a form of social constructionism (Hall, 1996).

2. Data

The article which produced the data for this study was published in the online version of *The Sun*, a British tabloid newspaper, on January 8th, 2013. The article, found in the News section, was titled, "*Thug breaks man's jaw outside takeaway in unprovoked attack...because he was ginger"* below which were two pictures taken from CCTV footage, the first showing a larger man punching another man. The second photograph shows the smaller individual falling to the ground in the street. After six sentences of the article, a CCTV video clip of the attack is embedded into the page. Further down, there is another picture which depicts the larger man exiting a store and confronting the smaller man and a fourth photograph which shows the moment in which the smaller man was hit.

The article states that a man was attacked and left seriously injured in what is described as *an unprovoked attacked*. The story contains a large proportion of direct quotes as the injured man describes the incident and the long period of physical and psychological recovery. He relates how he went into a pizza takeaway restaurant with his girlfriend and was verbally abused before being physically attacked upon leaving. He was left unconscious with a badly broken jaw and needed three months to recover. In the article he is clearly depicted as a blameless victim, whereas the other man is presented as the guilty aggressor. The article states that the attacker was still being sought by the police at the time of publication.

The article produced 190 responses from readers containing a total of 6,606 words. If readers wished to comment on the message board, they would first have to create an account by either using an existing Twitter or Facebook account, or by creating a new account with the newspaper. The reader would have to submit a user name; an avatar could also be added. Using these two sets of information, 58 posters used male names or provided pictures of males, whereas only 4 posters indicated that they were female. The other poster provided user names and avatars which did not indicate gender.

3. Methodology

Once the corpus of reader responses was compiled, the first stage of the analysis consisted of a study of frequency data. From the word list ordered by frequency, it was possible to gain an understanding of aspects of the corpus which occurred often and therefore had the potential to demonstrate the lexical choices that the tabloid readers who responded to the article made, which could relate to the presentation of particular discourses or attempts to construct identity.

Once a word list organised by frequency had been analysed, a word list arranged by keyness was observed. For this study, a reference corpus of general English was constructed which consisted of newspaper articles from the British newspaper *The Guardian*. Wordsmith Tools was set to perform a log likelihood statistical test for each word, which gave a probability value (p value). This value designates the degree of confidence that a word is key due to chance alone, the smaller the p value, the more likely that the word's presence in one of the corpora is not due to chance but the result of the author's choice to use the word consciously or subconsciously. Wordsmith uses a default of $p<0.000001$.

Once the keywords lists had been studied, to gain a more comprehensive insight into the data, the keywords were observed in context by undertaking a concordance analysis. A concordance-based study is able to disclose a range of discourses; therefore the notions of semantic preference and semantic prosody are important concepts. Semantic preference is

defined by Stubbs (2001: 65) as "the relation between a lemma or word form and a set of semantically related words", thus it is related to the notion of collocation. Semantic preference is the meaning which arises from the common semantic features of collocates of a given node word (McEnery et al. 2006: 84). Semantic preference is linked to the notion of semantic prosody (Louw, 1993) where patterns in discourse can be established between a word and a set of related words that indicate a discourse.

To gain further insights into the usage in context of the keywords, a study of collocates of those words was undertaken in which mutual information (MI) was utilised to calculate the strength of collocation. The MI score measures the degree of non-randomness present when two words co-occur. An MI score of 3 or higher can be considered to be significant (Hunston 2002: 71). However, one problematic issue with MI is that high scores may be achieved by relatively low frequency words, therefore this must be taken into consideration as words with a low frequency ought not to be considered as significant in spite of a high MI score.

Once this procedure had been completed using the whole corpus, a second corpus was compiled of texts by posters who indicated that they were male. This was achieved by observing the usernames or the avatars. If in each case the posters indicated that they were male, the text was added to the second corpora.

4. Findings

Frequency is one of the most fundamental concepts in the analysis of a corpus (Baker, 2006), which is able to provide insights which illuminate a range of themes. The twenty most frequent words in the corpus are as follows: *the* (289), *a* (203), *to* (176), *and* (152), *he* (117), *of* (103), *is* (102), *this* (90), *I* (86), *him* (84), *for* (78), *that* (74), *in* (69), *his* (59), *be* (58), *it* (53), *not* (53), *with* (52), *on* (49), *out* (49). It is apparent that the most frequent words in the corpus are grammatical words (function words). Such words belong to a closed grammatical class consisting of high frequency words such as articles, pronouns, conjunctions and prepositions. These groups of words do not necessarily provide insight to the discourses found within the corpus as most forms of language contain a high proportion of functional words. However, by taking into consideration the most frequent lexical words such as nouns, verbs, adjectives and lexical adverbs, a clearer notion of the discourses within the corpus is attained: *like* (41), *get* (40), *guy* (27), *victim* (25), *hope* (24), *someone* (24), *people* (23), *thug* (22), *ginger* (19), *punch* (19), *know* (18), *got* (17), *think* (17), *fight* (15), *man* (15), *prison* (15), *catch* (14), *scum* (14), *caught* (13), *coward* (13).

The second list presents a clearer picture of what the corpus is about. There are words associated with acts of violence (*victim, punch, fight*). Another group of words are used to describe a person negatively (*thug, scum, coward*). An additional aspect of the list is that *catch* and *caught* are both present, therefore the lemma CATCH is significant. When this is taken into consideration with *prison*, it can be seen that another theme is prominent. Beside the lemma CATCH, other verbs are also present (*like, get, hope, know, got, think*). Further analysis of collocational data and concordance lines will be needed to understand the context of these verbs within the corpus, although when the COCA corpus of general English (Davies, 2012) is referenced, it can be seen that such verbs are also of high frequency in a corpus of general English.

By considering frequency beyond the single word further insights can be gained. The most frequent 3-word clusters are: *is going to* (7), *he is a* (6), *lock him up* (6), *him up and* (5), *on the wrist* (5). By observing the concordance lines for the most frequent cluster *is going to*, certain themes can be seen: (1) *suspended, you just know He is going to get off lightly* (2) *moment of madness and this kid is going to pay the ultimate price* (3) *punching someone. This place is going to the dogs at a rapid rate* (4) *the lad who threw the punch is going to have to deal with the 'victim's friends* (5) *going to be mugged or someone is going to attack me, so im always ready* (6) *kick someones head in . Apparently UKIP is going to sort this out....!* (7) *just a stupid no brain thug who is going to jail.*

The writer of line 1 predicts the assailant is not going to be severely punished, whereas in lines 2 and 7 the opposite prediction is made. Line 4

remains within the theme of punishment, but states that the attacker is going to have to face the victim's friends. In line 3, the poster describes the decline of society and social behaviour, which is similar to the sentiment found in line 6, which states that a British political party claims to have a solution for such a situation, although the use of the word *apparently* appears to contest such a claim. The poster of line 5 describes the actions that he or she would take if attacked in similar circumstances, thus claiming to be more prepared to act in the instance of street violence than the victim was. Such examples highlight the diverse responses to the act of violence depicted in the article.

Of the six instances of *he is a,* five refer to the attacker. He is described as: *a self centred thug, a threat to the society* and *a trained fighter*. Although further analysis is necessary, the data demonstrates certain themes within the corpus; the assailant is condemned for his action and his fighting ability is discussed. Two clusters are evident which combine to produce the phrase: *lock him up and* followed by a phrase such as *throw away the key* demonstrate the punishment the posters consider the aggressor deserves. Another frequent cluster is part of the phrase: (*slap* or *smack*) *on the wrist,* thereby predicting that he will be dealt with lightly by the law. Thus it can again be seen, there are various reactions and opinions to the violence: the aggressor is condemned for his actions, that society is described as having poor moral standards, and that he will not receive adequate punishment for his actions.

It can be observed that by analysing frequency lists, discourses within the corpus may be highlighted. In the following section, the keywords of the corpus will be discussed.

The keywords with the highest levels of keyness are as follows: *him* (113.53), *this* (86.86), *I* (69.8), *guy* (66.26), *someone* (58.89), *hope* (58.89), *thug* (53.98), *victim* (53.56), *punch* (46.61), *ginger* (46.61), *get* (46.10), *he* (44.93), *like* (42.27), *why* (37.08), *your* (35.69), *catch* (34.34), *scum* (34.34), *don't* (31.88), *coward* (31.88), *out* (28.41). These words could be divided into three separate groups. There are functional words: *him, this, I, he, why, your, don't, out,* verbs: *hope, get, like, catch,* and nouns or adjectives: *guy, someone, thug, victim, punch, ginger, scum, coward*. It is not possible to present all the findings in this paper due to space restrictions, therefore certain keywords from each will group be selected for further analysis.

If the functional words are taken into consideration, *him* has the highest level of keyness and is the second most frequent among the keywords on the list. As this word is also most likely to be referring to one of the two male actors in this instance of street violence, further analysis may provide further insights as to how the two men are constructed. Therefore, the keyword *him* was studied in context by considering the concordance lines.

Of the 84 instances of *him,* 73 are referring to the attacker, of which 36 depict him negatively, 10 reference the victim and one refers to a person in a hypothetical situation, thus the aggressor appears to be the primary focus of the posters. If the collocates of *him* with the strongest levels of MI scores are calculated, the following list is provided: *suspended, teach, catch, example, sentence, years, really, throw, someone, prison*. This appears to indicate that within the corpus there is a dominant discourse associated with the attacker being caught and punished for his actions. This is confirmed when the word *him* is seen in context. A principle discourse focuses on the attacker being caught: *Catch him and jail him ASAP.* Another concordance line within the same semantic field describes the same notion more strongly: *Scum of Britain!!!PLEASE catch him*. Another example using a phrase which was found in the most frequent clusters is as follows: *Find that punk, lock him up and throw away the key*. Therefore it can be seen how the posters react to such acts of violence. Another discourse within the corpus denigrates the attacker, as the last example demonstrates with the term *punk*. Other examples include: *UK is full of scum like him! / No other word for him COWARD.*, and referring to him as a *mug brained idiot*. However, not all of the posters refer to the aggressor in such negative terms, nor do all the people who responded to the article believe that he should receive a prison sentence for his actions as the following examples indict: *but the other kid does square up to him / I doubt he wudda smacked him like that completely unprovoked / What if the 'victim' offered him out*

to begin with? This appears to indicate that there are some writers who do not accept the opinion of the article and are willing to consider alternative scenarios for the event which took place, thus indicating that the response to acts of violence among the message board posters is not homogenous.

Another keyword which provides insights into the corpus is *punch*. There are 19 instances of this word, all of which are in the form of a noun. When analysed in context, opposing discourses are evident; 9 of the lines either defend the attacker or are appreciative of his fighting skills, whereas only 5 instances denounce his actions. Another 5 instances of *punch* were classified as neutral, neither defending nor denouncing the attack. Examples of instances which depict the attack positively are as follows: *Boom! What a punch! / great punch,* and *the kid knows how to throw a punch*. Such examples are in contrast to the article which clearly denounced his actions. It can be seen that the writers of these examples value the act of violence regardless of the fact that it left one man seriously injured. As previously stated, the corpus does not contain a single discourse; other examples of *punch* in context are more condemning: *The guy who threw the punch is a bully / it was a dangerous cowardly sucker punch*. This brief study of *punch* demonstrates that both qualitative and quantitative analysis is necessary not only to discover discourses within a corpus, but also to comprehend their statistical significance.

Another keyword of interest is *victim*. There are 25 instances of this word; they all refer to the man who was left unconscious with a broken jaw. However, when the concordance lines are studied, it can be seen that a number of writers are using this word ironically when labelling this man as a victim of crime or cast doubt on the interpretation of events depicted by the newspaper. Seven of the writers do not consider the injured man to be blameless as the following examples demonstrate: *the person who hit the deck was not a victim / Looks like the 'victim' called the other guy out of the joint then got punched / I doubt that he's the complete victim he's made himself out to be*. These writers do not appear to accept the opinions of the newspaper nor the evidence provided by the link to the CCTV footage which clearly illustrates the assault. The writer of the second example places the word within quotation marks to emphasis the fact that it is doubted whether the person is in fact blameless. Other writers demonstrate a different opinion, as the following examples illustrate: *it looks like he had 20lbs over the victim / That victim could have been a brain op patient.*

4.1 Men Only Corpus

Once the analysis had been completed using the corpus of all the postings, a second corpus was compiled using posts in which the writers had indicated that they were male either via a user name or avatar. This procedure was undertaken in order to observe if the stating of gender had a significant impact on the findings. This corpus contained 2,088 tokens; the ten most frequent lexical words were as follows: *like* (14), *get* (12), *victim* (10), *punch* (9), *think* (9), *ginger* (8), *guy* (8), *looks* (8), *people* (8), *attack* (6). When this list is compared with the complete corpus, it can be seen that *punch* is much higher, as the posters discuss the blow which broke a man's jaw. Secondly, both *hope* and *thug* are no longer present. This appears to signify that those who indicated that they are men refer to the aggressor as a *thug* less frequently, nor do they use the word *hope* as often. When *hope* is studied in the first corpus, it is most often used in phrases whereby the writer expresses a wish that the aggressor is caught and sent to jail. The weakening of such a sentiment in the second corpus is of interest. This pattern of differences is reinforced when the keywords with the highest levels of keyness are observed: *him* (42.29), *punch* (38.79), *victim* (36.65), *this* (35.23), *ginger* (34.48), *guy* (34.48), *your* (32.37), *looks* (28.45), *I* (28.04), *coward* (25.85). *I* is now present on the list, which appears to indicate the posters who state that they are male express their opinions more frequently than those who do not. However, it must be noted that *coward* has a high level of keyness, reinforcing the notion that the posters are not a homogenous group with shared set of values.

When *punch* is observed in context, only one post refers to the assailant as a *bully*, the other instances of the word are used in phrases which

demonstrate an appreciation of his fighting technique or describe it in more detail. The posters use *victim* either ironically or by stating that he was not a victim, only a minority labelled the injured man as such. The word *ginger* also demonstrates the lack of support for the injured man in the corpus; of the eight instances of this word, six are used by posters when casting doubt on his degree of innocence in the act of violence. When *looks* is observed in context, again a significant difference from the first corpus is seen, as only one instance of this word is used in a context which criticises the attacker. Again the majority of posters cast doubts on the innocence of the second man as the following phrases demonstrate: *It looks to me like the ginger lad goes out for a fight. / This isn't as innocent as it looks. The ginger lad walks out first. / It looks like the ginger fella took offense.* Thus there appears to be less condemnation of the attacker in the second corpus than in the first.

The data demonstrates that the writers who posted on the newspaper message board in response to the article are clearly of differing opinions. There are those who accept the views presented by the newspaper which condemn the assailant and his actions, clearly articulating how he should be punished as a consequence of his actions. Others use the incident to express an opinion the England has and is still experiencing a decline in social standards and morality, and furthermore that law enforcement is too lenient to effectively respond to such a situation. However, there is another statistically significant discourse within the corpus which is contrary to those which condemn or criticise the violence. In this discourse, violence is seen as something which is appreciated and respected, where the aggressor is not depicted as the guilty party and where the victim is not seen as blameless. This second semantic field becomes much more evident when the data is grouped according to those who indicated that they were men.

5. Discussion

This corpus linguistic study, which is based on online responses to a newspaper article, contains sociological and cultural components. As the data was collected from a single newspaper, it is not possible to state that the findings reflect upon a larger social group other than those who posted on the website. Furthermore, the data does not provide insights into the level of influence the article had on the readers, as the stance taken by the newspaper journalist may, or may not have affected the responses found in the data. However, attempts were made not to decontextualise the data as the language of the article and images found both in the article and the avatars of the posters may have influenced certain stances taken by the writers. As the study was sociolinguistic in nature, the corpus was not annotated with a grammatical tagger, and due to the small size of data, it was possible to consider and review each post before deciding if it met the required criterion to be added the second corpus. The effect of irregular spelling found in the data was minimal, as again through the small size of the corpus, it was possible to return to the source of the data to observe the context in which it was located.

As previously stated, the data for this study is narrow in scope, and therefore does not shed light on stances held by individuals outside of the particular target group. However, the findings are of relevance in the fields of identity construction, masculinity and violence. The findings have demonstrated a continuum of opinions and stances on the online message board in response to a specific act of violence. Such expressions of opinions and stances may be considered to be a reflection of an aspect of identity the writer constructs for himself or herself. Early CMC scholars described how the Internet liberated people from social constrain through there being a supposedly unbiased and non-prejudiced environment. The Internet was also believed to provide a measure of anonymity; however this perception has now appeared to have lessened due to the rise of social media networks in which any form of Internet activity is traceable and where users are aware of a degree of accountability regardless of the spatial distance when interacting on the Internet (Thurlow *et al* 2004).

Only a very small number of posters indicated that they were female, in contrast to a much larger number who claimed to be male, although by using the information provided on the message board, a significant proportion of the posters

provided no information related to their gender. When the whole corpus was analysed, it appeared that the principle discourses within the texts were critical of the violence which took place. However, when the second corpus was built containing posts by writers who indicated that they were male, the discourse which accepted this violence and questioned the degree of innocence of the injured man, become much more prominent.

By manipulating the data in this manner, it has been possible to focus on the responses of individuals who stated masculinity to be part of their identity. Masculinity can be defined as the trait of behaving in ways that society considers to be typical and acceptable for males. Masculinity, like gender, is constructed and therefore is something that has to be worked at. Boys and men have to prove their masculinity constantly (Kimmel 2001: 269). One form of proving masculinity for certain individuals is to condone violence, as the data has shown. Hegemonic masculinities (Connell 1995) are characterised as the variety of masculinity capable of marginalising and dominating not only women, but also other men. It is dependent on subordinate masculinities, since it must contradict them. However, in the data presented in this paper, in can be seen that subordination is achieved through violence, where the weaker injured male has been subordinated and rejected by other men, although this form of action was not uniform, again demonstrating the lack of a homogenous response. Thus, one trait of hegemonic masculinity is the use and acceptance of violence against other men as a means to subordinate others.

For certain researchers, such as Whitehead (2002: 93-94), discourse is focused upon as a means to comprehend how men practice hegemonic masculinity and perform identity work. Masculine identities can therefore be understood as effects of discursive practices; they are fashioned within institutions and are historically constituted. An online message board such as the one used in this research is one location where identity may be constructed, practiced and maintained. One way that the gender order is maintained is by linking notions of appropriate and inappropriate gendered performances to different types of identities, and as the data has shown, one form of behavior which is seen to be acceptable by certain individuals, is the use of, or appreciation of violence.

According to Whitehead (2002: 33-34), the notion that masculinity is a singular rather than multiple identities has been viewed as problematic, particularly where gender identities and power relations are contextualised practices. In order to comprehend the diversity of masculinities, it is necessary to study the relations, such as subordination and dominance, between the different forms of masculinity. These relationships are constructed through practices that may intimidate or exploit others (Kiesling 2006: 118). The data has shown how these complex positions regarding responses to violence interact. Masculinity is not a fixed trait, but a social process dependent upon restatement, and which, in various forms, involves language, thereby centrally situating linguistic issues in the theorising of gender. Men who heavily invest in a particular masculinity will attempt to communicate in a manner particular for that specific trait (Moita-Lopes 2006: 294). Masculinities are not displaced from a social context, but embedded and implicated in the lives of men.

Responses to acts of violence may be considered by certain individuals to be a tool for both the creation of and defence of self-image. Using corpus linguistic methods, the data has highlighted discourses which demonstrate that a wide range of stances exist, which in turn signifies the plurality of masculinity. Analysis of the data has shown that there exists a subculture of violence, whereby acts of aggression are respected and esteemed, and therefore a means to construct a masculine identity.

Researchers such as Messerschmidt (2004) argue that with the loss of traditional industrial job opportunities and the shift towards a service-based economy, certain working class men have found new means of establishing masculinity; violence and street fights are one means of doing so. Such a form of masculinity emphasises toughness and a willingness to fight and defend oneself in the face of perceived threats or challenges by other males. The findings have shown that all men do not respond to violence in the same way. This will reflect upon an understanding of what men are and

the consequences of acts such as the one focused upon in this study. However, it has been shown that for certain men, violence is considered a means for validifying masculinity through peer support which encourages and legitimises acts of aggression. Hegemonic masculine discourses and practices, such as violence, may be learned through interactions, both virtual and face to face, which justify the relevance of studying online communication.

Although researchers such as Winlow (2001) consider violence amongst males to be a consequence of the destabilising effects of postmodernism, others such have Pinker (2011) describe how violence has been a constant trait throughout human history. Therefore, it may be argued that violence and aggression by men is more closely linked to aspects of patriarchal and hegemonic masculinity (Connell 1995) than it is with social responses to the effects of postmodernism.

6. Conclusion

The Internet is a location where individuals may construct identity by expressing stances and through interactions with other Internet users. Whenever an individual interacts in a social environment, an aspect of their identity is revealed, and as identity construction and maintenance is a continual process, further construction takes place; this also applies for an online environment. The identities that individuals construct and the interactions they make on locations such as message boards may not necessarily be totally reliable or accurate. However, Wiszniewski and Coyne (2002) argue that regardless of the reliability of the interaction or identity construction, a reflection of the authentic identity is formed which will reveal an aspect of the user's identity.

This paper has demonstrated that by employing a corpus linguistic approach, multiple expressions of identity and identity construction on the Internet may be studied. Discourses of violence and masculinities have been discussed, and the continuum of responses to violence observed and analysed. The results indicate that for certain individuals, violence and aggression are esteemed character traits, while others rejected and condemned them, thus confirming the notion of multiple masculine identity traits rather than a singular stereotypical construction. In addition, it has been shown that posters who state that they are men are more likely to regard interpersonal violence as an acceptable trait of masculinity.

Furthermore, the study has demonstrated that Web-derived data may be collected and filtered in various ways using contextual information in order to shed light on sociolinguistic and identity traits of particular target groups.

References

Anderson, E. (1990). *Streetwise: Race, Class and Change in an Urban Community*. Chicago: University of Chicago Press.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Bauman, Z. (2007). *Liquid Times.* Cambridge: Polity.

Benwell, B. & Stokoe, E. (2006). *Discourse and Identity*. Edinburgh: Edinburgh University Press.

Connell, R. W. (1995). *Masculinities*. Berkeley; Los Angeles, University of California Press.

Davies, M. (2012). COCA: Corpus Of Contemporary American English. http://corpus.byu.edu

DeKeseredy, W.S. & Schwartz, M.D. (2005). Masculinities and Interpersonal Violence. In M.S. Kimmel, J. Hearn, R.W. Connell (Eds.) *Handbook of Studies on Men & Masculinities.* Sage Publications: London (pp. 353-366).

Hall, J.K. (1996). Who needs "identity"? In S. Hall and P. du Gay (eds.) *Questions of cultural identity.* London: Sage, pp. 1-17.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Kiesling, S. F. (2006). Playing The Straight Man. In D. Cameron & D. Kulick (eds.) *The Language and Sexuality Reader*. Oxford, Routledge. pp. 118-131.

Kimmel, M. S. (2001). Masculinity as Homophobia: Fear, Shame and Silence in the construction of Gender Identity. In S. M. Whitehead & F. J. Barrett (eds.) *The Masculinities Reader*. Cambridge, Polity. pp. 266-287.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In Baker, M., Francis, G. & Tognini-Bonelli, E. (eds.) *Text and Technology: In honour of John Sinclair*. Philadelphia and Amsterdam: John Benjamins,157-176.

McEnery, T., Xiao, R. and Tono, Y. (2006). *Corpus-Based Language Studies: An advanced resource book*. London and New York: Routledge.

Messerschmidt, J. (2004). *Flesh and Blood: Adolescent Gender Diversity and Violence.* Lanham: Rowan and Little Field.

Moita-Lopes, L. P. (2006). On being white, heterosexual and male in a Brazilian school: multiple positionings in oral narratives. In A. de Fina, D. Schiffrin & M. Bamberg (eds.) *Discourse and Identity*. Cambridge: Cambridge University Press.

Pinker, S. (2011). *The Better Angels of Our Nature: Why Violence Has Declined.* New York: Viking.

Stubbs, M. (2001). *Words and Phrases*. Oxford: Blackwell.

Thurlow, C., Lengel, L. & Tomic, A. (2004). *Computer Mediated Communication: Social Interaction and the Internet.* Thousand Oaks, CA: Sage Publications.

Wiszniewski, D. & Coyne, R. (2002). Mask and Identity: The Hermeneutics of Self-Construction in the Information Age. In K. A. Renninger & W. Shumar (eds.) *Building Virtual Communities* (pp. 191-214). New York, New York: Cambridge Press.

Whitehead, S. M. (2002). Men and Masculinities: key themes and new directions. Cambridge, Polity Press.

Winlow, S. (2001). *Badfellas: Crime, Tradition and New Masculinities.* Berg: Oxford.

A web-based model of semantic relatedness and the analysis of electro-encephalographic (EEG) data

Colleen E Crangle
School of Computing and Communications
Lancaster University, UK
Converspeech LLC, Palo Alto, CA, USA
crangle@converspeech.com

Patrick Suppes
Center for the Study of Language and Information
Stanford University, CA, USA
psuppes@stanford.edu

Abstract

Recent studies of language and the brain have shown that models of semantics extracted from web-based corpora can predict brain activity. This paper shows how a model of semantic relatedness extracted from the web can predict the brain activity for relations between words. The model uses ukWaC, a large corpus of English obtained from the web, along with co-occurrence frequencies and mutual information scores, to represent the strength of associations between words. Brain data obtained from participants while they are assessing the truth or falsity of English-language statements provide a model of the associations between words as perceived by the participants. The brain-data model and the semantic model are compared and the strength of similarity between the two is assessed. Corpus-based studies of semantics and the brain potentially offer a new way to validate any proposed corpus-based model of semantics.

1 Introduction

Recent studies of language and the brain have shown that models of semantics extracted from web-based corpora can predict brain activity as measured by functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG), or electroencephalography (EEG). In this paper, we show that a model of semantic relatedness extracted from the web can predict the brain activity for the relations between words. Unlike previous work in which semantic features of individual words are predicted, our work examines sets of words and predicts the relations between them. The brain data are drawn from experiments in which statements about commonly known geographic facts of Europe were presented auditorily to participants who were asked to determine the truth or falsity of each statement while EEG recordings were made (Suppes *et al.*, 1999; Suppes *et al.*, 2009). For the semantic model we use ukWaC, a large (> 2 billion token) corpus of English constructed by crawling the.uk Internet domain (Baroni *et al.*, 2009), along with co-occurrence frequencies combined with pointwise mutual information (Turney, 2001). The aim is to estimate the strength of the association between any two words and use those estimates to construct a network of relations for a given set of words. That network can then be compared to the network of relations we find in brain data.

2 Background

Crangle *et al.* (2013) describes a method by which structural similarities between brain data and linguistic data at the semantic level can be assessed. It further shows how to measure the strength of these structural similarities and so determine the relatively better fit of brain data with one semantic model over another. Two semantic models were investigated, WordNet (Fellbaum, 1998) and latent semantic analysis (LSA, Landauer and Dumais, 1997), with WordNet clearly emerging as the better predictor of brain activity. However, the rich resources of the Web hold obvious appeal as a source of semantic information and in this paper we examine the extent to which a web-derived model of semantics can predict brain activity for a set of words and the relations between them.

In other studies where web-based corpora have been used to predict brain activity, the semantics of a word has been represented by its distributional properties in the corpus, with the prediction limited to the category (tool or mammal, for example) of the word the participant is attending to. In Mitchell *et al.* (2008), for example, the semantics of a word was given by its distributional properties in the Google Inc. dataset consisting of English word n-grams and their frequencies (Brants and Franz, 2006). Taking 60 nouns referring to physical objects, the semantics of these nouns was given by their co-occurrence patterns with 25 manually-selected sensory-

motor verbs from the approximately 1-trillion-word set of web pages. Statistically significant predictions were made as to the semantic category (mammal or tool) of the words in this set using fMRI images collected while participants were attending to the words one by one.

Since Mitchell, other web-based corpora and other ways of selecting semantic features have been investigated to see if they offered improved methods of predicting from brain data the semantics of a word someone is seeing or hearing or otherwise attending to. Murphy *et al.* (2012) used a 16-billion-word set of English-language web-page documents and point-wise mutual information with co-occurrence frequencies to provide a category-focused semantic model. Pereira *et al.* (2010) used a large text corpus consisting of pertinent articles from Wikipedia and latent Dirichlet allocation (LDA, Blei *et al.*, 2003). Jelodor *et al.* (2010) took WordNet as a supplementary source of information, using WordNet's similarity measures instead of co-occurrence statistics to measure the strength of the association between the 60 nouns and the 25 sensory-motor verbs of Mitchell.

This paper represents the semantics of a word in terms of its place in a network of words, with the links in the network capturing the strength of the associations between any two words relative to the strength of the association between each of those words and all other words in a set of words. This network of association strengths is then compared to the network of associations revealed by the brain data. The next section shows how the network of relations is constructed from ukWaC, using the 50% randomized subset of ukWaC available through CQPweb (Hardie, 2012) along with collocates and the measure of mutual information provided by the CQPweb interface. It then shows the method by which the brain and semantic data are compared.

3 Method and Materials

3.1 Brain data

Brain data were derived from an experiment in which 48 spoken sentences about the geography of Europe were presented to nine participants (S10, S24, S25, S27, S16, S26, S12, S13, S18) in 10 randomized blocks, with all 48 sentences occurring once in each block. Half of the sentences were true, half false, half positive, and half negative (e.g., *The capital of Italy is Paris* and *Paris is not east of Berlin*). The possible forms of these sentences are: *X is [not]W of Y, W of Y is [not]X, X is [not]Z of X, Y is [not] Z of Y*, where $X \in$ {*Berlin, London, Moscow, Paris, Rome, Warsaw, Madrid, Vienna, Athens*), $Y \in$ {*France, Germany, Italy, Poland, Russia, Austria, Greece, Spain*), $W \in$ {*the capital, the largest city*}, $Z \in$ {*north, south, east, west*}, and *[not]* indicates the optional presence of *not*. The 21 country names, city names, and the four directions or relative locations are the words of interest for our analysis.

Since sentences are understood incrementally, word-by-word, brain activity time-locked to the presentation of the words in the 480 trials each participant is presented with can be analyzed to tell us something about language processing. Segments of the brain-wave data for each of the 21 words of interest were therefore extracted from each of the 480 trials presented to each participant.

3.2 Semantic data

UKWAC50 is a randomized subset of ukWaC in which each of the web pages sampled in ukWaC had a 50% chance of being included. It consists of 1,346,675 texts containing 1,127,056,026 words. For each word ω_i from our set of words $W = \{\omega_1, \omega_2, ..., \omega_{21}\}$ we found the case-sensitive collocates of ω_i with each of the other members of W using a window of 10 words before and 10 after ω_i, and with a minimum frequency of 5 for the occurrences of ω_i and the collocates. For each such collocate we took its mutual information score and used those scores to construct a 21-by-21 matrix $Q = (q_{ij})$, where each q_{ij} is the mutual information score for ω_i and ω_j normalized to the [0,1] interval by subtracting the overall minimum and dividing by the overall maximum minus the overall minimum, and assigning the maximum score of 1 on the diagonal to each word's association with itself. See Figure A1 in Appendix A.

3.3 Approach

We used 10 words at a time to compare the brain and semantic data, selecting three city names, three country names, and the four relative location words. Each such set of 10 words was then investigated to find out what the brain data revealed about the participants' perceptions of the relations between the 10 words. These perceptions were then compared to the semantic representation of the relations between the 10 words. Results for the following sets of 10 words are

reported in this paper: {*London, Moscow, Paris, north, south, east, west, Germany, Poland, Russia*}; {*Paris, Vienna, Athens, north, south, east, west, Italy, Spain, Austria*}; and {*Berlin, Rome, Warsaw, north, south, east, west, France, Greece, Poland*}.

3.4 Analyzing the brain data

The first step in analyzing the brain data is to use a statistical model to predict to which class a brain-data sample belongs, where each class corresponds to one of the 10 words in the set of interest. In order to classify the segments of data obtained from the individual trials, we use a linear discriminant model in a 5-fold cross-validation loop. The dimensionality of the data is reduced using a nested principal component analysis after ocular artifacts are removed from the brain-data samples using blind source separation. The procedure is described in detail in Perreau-Guimaraes *et al.* (2007).

For each group of 10 words, each brain-data sample is classified as a representation of one of the 10 words, that is, each sample is predicted to belong to one of the 10 words. For the set of words{*London, Moscow, Paris, north, south, east, west, Germany, Poland, Russia*}, for example, there are 640 data samples that are classified into 10 classes.

More generally, T brain-data samples s_1, s_2, ..., s_T are classified into the N classes ω_1, ω_2, ..., ω_N. If test sample s_i is classified as ω_i then s_i and ω_i have a minimal similarity difference compared to the other possible classifications.

Let $M = (m_{ij})$ be the confusion matrix for a given classification task, where m_{ij} is the number of test samples from class ω_i classified as belonging to class ω_j. By computing the relative frequencies $m_{ij} / \sum_j m_{ij}$ we obtain N-by-N estimates for the conditional probability densities that a randomly chosen test sample from class ω_i will be classified as belonging to class ω_j. Figure 1 gives the conditional probability estimates computed from the confusion matrix resulting from the classification of the brain data for the set of 10 words{*London, Moscow, Paris, north, south, east, west, Germany, Poland, Russia*} for participant S18. Following the conventions for heat maps, the higher the value of each element in the matrix the darker its shading.

	London	Moscow	Paris	north	south	east	west	Germany	Poland	Russia
London	0.275	0.108	0.133	0.042	0	0.008	0	0.075	0.025	0.033
Moscow	0.133	0.408	0.167	0.008	0.025	0.025	0.008	0.025	0.025	0.033
Paris	0.05	0.158	0.208	0.05	0.017	0.05	0.033	0.05	0.008	0.033
north	0	0.033	0.067	0.617	0.017	0	0.017	0.05	0.017	0.033
south	0.05	0.033	0.017	0	0.367	0.05	0.083	0.067	0	0.067
east	0.033	0	0	0.1	0	0.517	0.283	0.017	0	0
west	0	0.033	0.05	0.017	0.017	0.25	0.433	0.017	0.017	0.017
Germany	0.167	0.15	0.1	0.05	0.05	0.033	0	0.217	0	0.1
Poland	0.017	0.017	0	0.033	0.017	0.017	0	0.017	0.6	0.017
Russia	0.017	0.083	0.083	0.033	0.017	0	0.017	0	0	0.6

Figure 1: Conditional probability density estimates (shown as a heat map) computed from the confusion matrix resulting from the classification of 640 brain wave samples from participant S18 for the set of words {*London, Moscow, Paris, north, south, east, west, Germany, Poland, Russia*}.

Let these conditional probability estimates be given by the matrix $P = (p_{ij})$. **For each class ω_i** we then define a quaternary relation R such that $\omega_i \omega_j R \omega_i \omega_k$ if and only if $p_{ij} < p_{ik}$, that is, if and only if the probability that a randomly chosen sample from class ω_j will be classified as belonging to class ω_i is smaller than the probability that a randomly chosen test sample from class ω_k will be classified as belonging to class ω_i. R is an ordinal relation of similarity differences, a partial order that is irreflexive, asymmetric, and transitive.[1] [2]

We then form the relational structure (W, R) which is constructed from the N partial orders R (one for each ω_i) and the finite set W of classes ω_i together with the real-valued functions given by the inequalities $p_{ij} < p_{ik}$. This relational structure (W, R) provides a formal characterization of the brain data that captures the relations between the elements of W.

3.5 Further characterization of the semantic data

For a given set of N words $W = \omega_1, \omega_2, ..., \omega_N$ an N-by-N matrix $Q' = (q'_{ij})$ is compiled from the 21-by-21 matrix Q described earlier of normalized mutual information scores derived from ukWaC. Figure 2 shows the association matrix for the set of 10 words{*London, Moscow, Paris, north, south, east, west, Germany, Poland, Russia*}.

[1] Specifically, the function defined by these inequalities is the function f defined on A such that xy R uv iff f(x) - f (y) < f (u) - f (v), with f(x) - f(y) represented by p_{xy}.

[2] Note that an ordinal relation of similarity differences, a partial order that is irreflexive, asymmetric, and transitive, is defined for each class ω_i. We therefore have N partial orders and the notation (W, R) will from here on be understood to refer to a structure with N partial orders defined on the set W.

	London	Moscow	Paris	north	south	east	west	Germany	Poland	Russia
London	1	0.292	0.4119	0.3563	0.3989	0.4255	0.3683	0.1153	0.1699	0.123
Moscow	0.292	1	0.5519	0.1762	0.1965	0.2231	0.1548	0.316	0.422	0.7815
Paris	0.4119	0.5519	1	0.1212	0.1687	0.1318	0.082	0.3396	0.3049	0.294
north	0.3563	0.1762	0.1212	1	0.5814	0.5624	0.5542	0.2266	0.182	0.2882
south	0.3989	0.1965	0.1687	0.5814	1	0.5757	0.5633	0.3565	0.3091	0.3162
east	0.4255	0.2231	0.1318	0.5624	0.5757	1	0.5419	0.1851	0.3052	0.3219
west	0.3683	0.1548	0.082	0.5542	0.5633	0.5419	1	0.2076	0.2088	0.2777
Germany	0.1153	0.316	0.3396	0.2266	0.3565	0.1851	0.2076	1	0.6927	0.599
Poland	0.1699	0.422	0.3049	0.182	0.3091	0.3052	0.2088	0.6927	1	0.7224
Russia	0.123	0.7815	0.294	0.2882	0.3162	0.3219	0.2777	0.599	0.7224	1

Figure 2: Association matrix (shown as a heat map) derived from UKWAC50 using collocates and mutual information scores for the set of words {*London, Moscow, Paris, north, south, east, west, Germany, Poland, Russia*}.

Then, as we did for the conditional probability estimates derived from the brain data, **for each class ω_i** we define a quaternary relation R' such that $\omega_i\,\omega_j\,R'\,\omega_i\,\omega_k$ if and only if $q'_{ij} < q'_{ik}$, that is, if and only if the difference between the association scores for ω_i and ω_j is smaller than the difference between the association scores for ω_i and ω_k. R' is an ordinal relation of similarity differences, a partial order that is irreflexive, asymmetric, and transitive.

We then form the relational structure (W, R') constructed from the N partial orders R' (one for each ω_i) and the finite set W of classes ω_i together with the real-valued functions given by the inequalities $q'_{ij} < q'_{ik}$. This relational structure provides a formal characterization of the linguistic data, one that captures the semantic relations between the elements of W.

3.6 Comparing brain and semantic data

Isomorphism between (W, R) and (W, R') would constitute the strongest measure of structural similarity between the brain and linguistic data, but isomorphism is almost certainly too strong a requirement for brain data obtained under current experimental conditions. Instead, we use a generalization of isomorphism, namely the notion of *invariant partial order*. For each class ω_i, we take R (partial order derived from the brain data) and R' (partial order derived from the linguistic data) and we compute their Spearman rank correlation coefficient. Those that have a statistically significant correlation are selected and their intersection is calculated. This intersection is that part of R and R' that is invariant with respect o each other. The intersection of two partial orders is also a partial order and so we have an invariant partial order for those ω_i for which the brain and the linguistic data are sufficiently strongly correlated. The number of such invariant partial orders gives a measure of the strength of the structural similarity between the brain and linguistic data for a given set W of words relative to a given semantic model.

To illustrate the formation of invariant partial orders we show in Table 1 two partial orders R and R' relative to ω_i = *London*. The brain data (R) is taken from Figure 1 and the linguistic data (R') is taken from elsewhere for illustrative purposes.

Linguistic data		Brain data	
R'		R	
London	1.000	*London*	0.275
Paris	0.466	*Paris*	0.133
Moscow	0.396	*Moscow*	0.108
Germany	0.322	*Germany*	0.075
Russia	0.303	*north*	0.042
Poland	0.299	*Russia*	0.033
north	0.106	*Poland*	0.025
south	0.103	*east*	0.008
west	0.078	*south*	0.00
east	0.076	*west*	0.000

Table 1: Two partial orders R and R' relative to ω_i = *London*

Figure 3 gives a graphical representation of the invariant partial order for the two partial orders of Table 1.

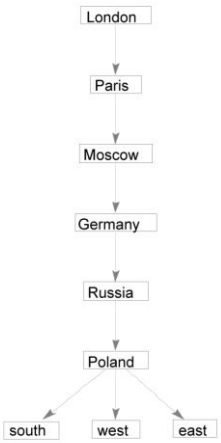


Figure 3: Graphical representation of the invariant partial order for the two partial orders of Table 1

3.7 Summary of method

Using brain-data samples from individual trials time-locked to the presentation of each word of interest, partial orders of similarity differences are computed for the brain data. For the linguistic data, partial orders of similarity differences are computed from the mutual-information collocation scores derived from ukWaC. Those relations that are invariant with respect to the brain and linguistic data, and are correlated with sufficient statistical strength, amount to structural similarities between the brain and linguistic data. The number of such invariant partial orders gives a measure of the strength of the structural similarity.

For each of the nine participants we computed 30 single-trial classifications of the brain data (using random resampling with replacement) for each set of 10 words given in section 3.3 and we took the average of the confusion matrices to compute an estimate of the conditional probability estimates. We also computed the association matrix derived from ukWaC for each of these sets of words. We then found for each set of 10 words the partial orders that were significantly highly correlated (rho=.6485, $p < 0.05$) and invariant with respect to the linguistic data and the brain data for each participant. The total number of such invariant partial orders represents the strength of the structural similarity between the brain data and the ukWaC -derived semantic data.

We further computed association matrices for the same three sets of 10 words using LSA and found the total number of significant invariant partial orders relative to this semantic model for each of the participants. We used the application at http://lsa.colorado.edu/ (accessed January 15, 2013). The computations were based on texts of general reading up to 1st year college level and a maximum of 300 factors was permitted in the analysis.

4 Results

For each of the nine participants, the number of invariant partial orders computed using the ukWaC-derived semantic model was greater than that computed using LSA. Figure 4 contains the results.

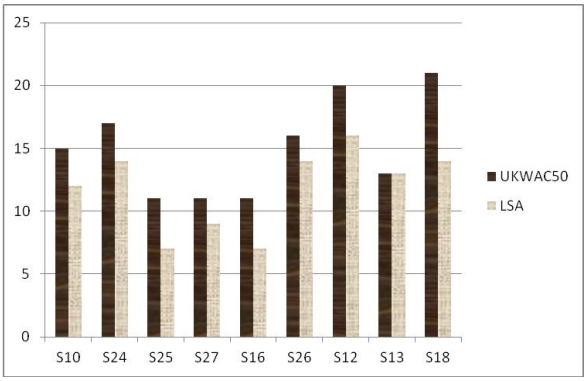


Figure 4: The number of invariant partial orders computed using the UKWAC50 corpus (along with collocates and mutual information) compared with the number of invariant partial orders computed using LSA

5 Discussion

The relatively better fit of the ukWaC-derived semantic model over the LSA-derived model suggests that it (the corpus and the method of deriving distributional properties from it) is a good match to the semantic knowledge of language users. Further investigations using different corpora and/or different methods of assessing the strength of the association between words could give greater insight into how semantic information is represented in the brain.

One obvious area for investigation concerns the use of the mutual information score. This score reflects the extent to which observed frequencies of co-occurrence differ from chance, and in this it represents a measure of the strength of association between two words. Its calculation uses the number of times the words appear together versus the number of times the words appear separately. The mutual information score can be unduly high for low-frequency words. It will promote any pair of words for which the frequency of co-occurrence is high relative to the frequency of occurrence of either of the two words. For example, the name of Haile Selassie (Emperor of Ethiopia from 1930 to 1974) will score high because the chances of one word appearing without the other are low. The t-score provides a way of mitigating this potential problem. The t-score promotes co-occurrences that are well attested, that is co-occurrences for which there have been a reasonable number of appearances. In these cases the strength of association given by mutual information may not be high but the confidence that there is some association, as measured by the t-score, is high. A combination of t-score and mutual information may give best results, that is, may better represent the word associations we find in brain data.

Since Lund and Burgess (1996) distributional corpus-based models, ranging from simple co-occurrence vectors to probabilistic topic-based approaches such as LDA, have increasingly gained acceptance. One important point about distributional models is that they require validation. Typically, validation entails seeing how well a given model correlates with the semantic judgments of native speakers. When speakers' predictions diverge from a model, questions arise as to whether that divergence points to limitations in the model or inadequacies in how those judgments were solicited, whether they were fine-grained enough, for example. Corpus-based studies of semantics and the brain potentially offer a new way to answer these questions.

References

Baroni M, Bernardini S, Ferraresi A and Zanchetta E (2009) The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation 43 (3): 209-226.

Blei, David M., Andrew Y. Ng , Michael I. Jordan, Latent dirichlet allocation, The Journal of Machine Learning Research, 3, p.993-1022, 3/1/2003

Brants, T. and Franz, A. (2006). Web 1T 5-gram Version 1.

Crangle CE, Perreau-Guimaraes M, Suppes P (2013) Structural Similarities between Brain and Linguistic Data Provide Evidence of Semantic Relations in the Brain. PLoS ONE 8(6): e65366. doi:10.1371/journal.pone.0065366.

Fellbaum C (Ed.) (1998) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Hardie, A (2012) "CQPweb - combining power, flexibility and usability in a corpus analysis tool". International Journal of Corpus Linguistics 17(3): 380-409. http://dx.doi.org/10.1075/ijcl.17.3.04har

Jelodar, A B, M Alizadeh, S Khadivi (2010) WordNet Based Features for Predicting Brain Activity associated with meanings of noun. Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics, June 2010, Los Angeles, USA, Association for Computational Linguistics, 18—26

Landauer TK, Dumais ST (1997) A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review 104: 211-240.

Lund K, Burgess C (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instrumentation, and Computers 28: 203-208.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting Human Brain Activity Associated with the Meanings of Nouns. Science, 320:1191--1195.

Murphy, Brian, Partha Talukdar, and Tom Mitchell. 2012. Selecting corpus-semantic models for neurolinguistic decoding. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*(SemEval '12). Association for Computational Linguistics, Stroudsburg, PA, USA, 114-123.

Pereira F, M Botvinick, G Detre (2010) Learning semantic features for fMRI data from definitional text. Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics, June 2010,Los Angeles, USA, Association for Computational Linguistics, 1—9, http://www.aclweb.org/anthology/W10-0601

Perreau-Guimaraes M, Wong DK, Uy ET, Grosenick L, Suppes P (2007) Single-trial classification of MEG recordings. IEEE Transactions on Biomedical Engineering 54: 436–443.

Suppes P, Han B, Epelboim J, Lu Z-L (1999) Invariance between subjects of brain wave representations of language. Proceedings of the National Academy of Sciences, 96, 12953–12958.

Suppes P, Perreau-Guimaraes M, Wong DK (2009) Partial Orders of Similarity Differences Invariant Between EEG-Recorded Brain and Perceptual Representations of Language. Neural Computation 21, 3228–3269.

Turney P (2001) Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In L. De Raedt & P. Flach (Eds.), Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001) (pp. 491-502). Freiburg, Germany.

Appendix A: Matrix of web-derived associative scores

	Berlin	London	Moscow	Paris	Rome	Warsaw	Madrid	Vienna	Athens	France	Germany	Italy	Poland	Russia	Austria	Greece	Spain	north	south	east	west
Berlin	1	0.328	0.604	0.649	0.536	0.649	0.603	0.749	0.449	0.331	0.683	0.39	0.436	0.364	0.432	0.344	0.241	0.122	0.087	0.393	0.343
London	0.328	1	0.292	0.412	0.236	0.251	0.349	0.289	0.209	0.141	0.115	0.132	0.17	0.123	0.02	0.123	0.1	0.356	0.399	0.426	0.368
Moscow	0.604	0.292	1	0.552	0.43	0.636	0.567	0.519	0.377	0.176	0.316	0.287	0.422	0.781	0.207	0.331	0.257	0.176	0.197	0.223	0.155
Paris	0.649	0.412	0.552	1	0.588	0.541	0.652	0.614	0.372	0.649	0.34	0.378	0.305	0.294	0.221	0.312	0.311	0.121	0.169	0.132	0.082
Rome	0.536	0.236	0.43	0.588	1	0.517	0.582	0.555	0.49	0.286	0.238	0.73	0.271	0.198	0.29	0.755	0.376	0.181	0.207	0.214	0.131
Warsaw	0.649	0.251	0.636	0.541	0.517	1	0.603	0.721	0.497	0.259	0.367	0.338	1	0.462	0.422	0.158	0.287	0.117	0.153	0.244	0.165
Madrid	0.603	0.349	0.567	0.652	0.582	0.603	1	0.561	0.425	0.298	0.288	0.355	0.29	0.179	0.367	0.114	0.795	0.185	0.142	0.155	0.093
Vienna	0.749	0.289	0.519	0.614	0.555	0.721	0.561	1	0.391	0.272	0.402	0.425	0.392	0.291	0.936	0.227	0.248	0.095	0.131	0.152	0.095
Athens	0.449	0.209	0.377	0.372	0.49	0.497	0.425	0.391	1	0.081	0.151	0.266	0.153	0.049	0.172	0.8	0.166	0.14	0.052	0.026	0
France	0.331	0.141	0.176	0.649	0.286	0.259	0.298	0.272	0.081	1	0.739	0.736	0.565	0.579	0.714	0.679	0.695	0.227	0.357	0.185	0.208
Germany	0.683	0.115	0.316	0.34	0.238	0.367	0.288	0.402	0.151	0.739	1	0.74	0.693	0.599	0.805	0.678	0.646	0.227	0.357	0.185	0.208
Italy	0.39	0.132	0.287	0.378	0.73	0.338	0.355	0.425	0.266	0.736	0.74	1	0.664	0.531	0.773	0.805	0.763	0.277	0.298	0.205	0.169
Poland	0.436	0.17	0.422	0.305	0.271	1	0.29	0.392	0.153	0.565	0.693	0.664	1	0.722	0.73	0.647	0.661	0.182	0.309	0.305	0.209
Russia	0.364	0.123	0.781	0.294	0.198	0.462	0.179	0.291	0.049	0.579	0.599	0.531	0.722	1	0.62	0.527	0.487	0.288	0.316	0.322	0.278
Austria	0.432	0.02	0.207	0.221	0.29	0.422	0.367	0.936	0.172	0.714	0.805	0.773	0.73	0.62	1	0.74	0.661	0.163	0.257	0.189	0.177
Greece	0.344	0.123	0.331	0.312	0.755	0.158	0.114	0.227	0.8	0.679	0.678	0.805	0.647	0.527	0.74	1	0.754	0.754	0.262	0.273	0.188
Spain	0.241	0.1	0.257	0.311	0.376	0.287	0.795	0.248	0.166	0.695	0.646	0.763	0.661	0.487	0.661	0.754	1	0.269	0.353	0.207	0.188
north	0.122	0.356	0.176	0.121	0.181	0.117	0.185	0.095	0.14	0.227	0.227	0.277	0.182	0.288	0.163	0.754	0.269	1	0.581	0.562	0.554
south	0.087	0.399	0.197	0.169	0.207	0.153	0.142	0.131	0.052	0.357	0.357	0.298	0.309	0.316	0.257	0.262	0.353	0.581	1	0.576	0.563
east	0.393	0.426	0.223	0.132	0.214	0.244	0.155	0.152	0.026	0.185	0.185	0.205	0.305	0.322	0.189	0.273	0.207	0.562	0.576	1	0.542
west	0.343	0.368	0.155	0.082	0.131	0.165	0.093	0.095	0	0.208	0.208	0.169	0.209	0.278	0.177	0.188	0.188	0.554	0.563	0.542	1

Figure A1: Matrix of UKWAC50-based association scores derived using collocates and mutual information