Adrien Barbaresi, Kay-Michael Würzner

For a fistful of blogs: Discovery and comparative benchmarking of republishable German content

In: Faaß, Gertrud / Ruppenhofer, Josef (eds.): WORKSHOP PROCEEDINGS OF THE 12TH EDITION OF THE KONVENS CONFERENCE. Hildesheim: 2014, S. 2-10

Persistent Identifier: urn:nbn:de:kobv:b4-opus-26276

For a fistful of blogs: Discovery and comparative benchmarking of republishable German content

Adrien Barbaresi
Berlin-Brandenburgische
Akademie der Wissenschaften
barbaresi@bbaw.de

Kay-Michael Würzner
Berlin-Brandenburgische
Akademie der Wissenschaften
wuerzner@bbaw.de

Abstract

We introduce two corpora gathered on the web and related to computer-mediated communication: blog posts and blog comments. In order to build such corpora, we addressed following issues: website discovery and crawling, content extraction constraints, and text quality assessment. The blogs were manually classified as to their license and content type. Our results show that it is possible to find blogs in German under Creative Commons license, and that it is possible to perform text extraction and linguistic annotation efficiently enough to allow for a comparison with more traditional text types such as newspaper corpora and subtitles. The comparison gives insights on distributional properties of the processed web texts on token and type level. For example, quantitative analysis reveals that blog posts are close to written language, while comments are slightly closer to spoken language.

1 Introduction

1.1 Corpora from the web and CMC corpora

Web corpora can be useful to explore text types or genres which are not found in traditional corpora, as well as a whole range of user-generated content and latest language evolutions. The main issues when dealing with such web corpora, be it general-purpose corpora or specific ones, include the discovery of linguistically relevant web documents, the removal of uninteresting parts (or noise), the extraction of text and metadata, and last the republishing of at least part of the content.

So far, there are few projects dealing with computer-mediated communication. In the case of German, the DeRiK project (*Deutsches Referenzkorpus internetbasierte Kommunikation*) features ongoing work with the purpose to build a reference corpus dedicated to computer-mediated communication (Beißwenger et al., 2013).

More specifically, this kind of corpus can used to find relevant examples for lexicography and dictionary building projects, and/or to test linguistic annotation chains for robustness. The DWDS lexicography project at the Berlin-Brandenburg Academy of Sciences already features a good coverage of specific written text genres such as newspaper articles (Geyken, 2007). We wish to conduct further experiments including Internet-based text genres.

1.2 Problems to solve

The problems to solve in order to be able to build reliable computer-mediated communication (CMC) corpora are closely related to the ones encountered when dealing with general web corpora and described above. Specific issues are threefold. First, what is relevant content and where is it to be found? Second, how can information extraction issues be tackled? Last, is it possible to get a reasonable image of the result in terms of text quality and diversity?

Problem 1: Website discovery

First of all, where does one find "German as spoken/written on the web"? Does it even concretely exist or is it rather a continuum? Considering the ongoing shift from web *as* corpus to web *for* corpus, mostly due to an expanding web universe and the potential need for a better text quality, it is obvious that only a small portion of the German web space is to be explored.

Now, it is believed that the plausible distributions of links between hosts follows a power law (Biemann et al., 2013). By way of consequence, one may think of the web graph as a polynuclear structure where the nuclei are quite dense and well-interlinked, with a vast, scattered periphery and probably not so many intermediate pages somewhere in-between. This structure has a tremendous impact on certain crawling strategies. There are ways to analyze these phenomena and to cope with them (Barbaresi, 2014a), the problem being that there are probably different linguistic realities behind link distribution phenomena. While these notions of web science may seem abstract, the centrality and weight of a website could be compared to the difference between the language variant of the public speaker of an organization, and the variants among its basis.

Problem 2: Content extraction

Content extraction is a real problem concerning large web corpora (Schäfer et al., 2013), e.g. because of exotic markup and text genres. While it is generally possible to filter out tag clouds, post lists and left/right columns on webpage scale, the lack of metadata in "one size fits all" web corpora may still undermine the relevance of web texts for linguistic purposes.

In fact, one may argue that decent metadata extraction is necessary for the corpora to become scientific objects, as science needs an agreed scheme for identifying and registering research data (Sampson, 2000).

Problem 3: Text quality

In our particular context, we understand text quality in terms of usefulness for linguistic research. This type of quality has much to do with text integrity, cleaning, and preprocessing, and only addresses to a lesser extent intrinsic factors such as subtlety of language. Our approach deals with opening "black box corpora" and putting them on a test bench.

Undoubtedly, quality of content extraction has an effect on text quality, since the presence of boilerplate (HTML code and superfluous text) or the absence of significant text segments hinder linguistic work. Moreover, there are intrinsic factors speaking against web texts, for instance machine-generated and/or machine-translated content which leads to fluency and grammar correctness problems (Arase and Zhou, 2013), or mixed-language documents (King and Abney, 2013).

In sum, naive approaches to web crawling and web texts may yield positive results when text quantity is more important than text quality, e.g. in machine translation (Smith et al., 2013), but they are bound to impede proper linguistic research. In fact, there are (corpus) linguists who advocate a meticulous selection and extraction of web texts, since size cannot necessarily compensate for lack of quality (Biemann et al., 2013).

Possible ways to address aforementioned problems

We present three possible ways to cope with the issues described in this section. First, design an intelligent crawler targeting specific content types and platforms in order to allow for a fruitful website discovery and, second, to allow for the crafting of special crawling and content extraction tools. Third, find metrics to compare Internet-based resources with already known, established corpora, and assess their suitability for linguistic studies.

2 Retrieval of blog posts and corpus building

2.1 Blog discovery on *wordpress.com*

We chose a specific blogging software, WordPress, and targeted mostly its platform, because this solution compared favorably to other platforms and software in terms of blog number and interoperability. First, *wordpress.com* contains potentially more than 1,350,000 blogs in German. Second, extraction procedures on this website are

https://wordpress.org/
http://wordpress.com/stats

transferable to a whole range of self-hosted websites using WordPress, allowing to reach various blogger profiles thanks to a comparable if not identical content structure.

The crawl of the wordpress.com website has been prepared by regular visits of a tags homepage listing tags frequent used in German posts. Then, a crawl of the tag pages enabled us to collect blog URLs as well as further tags. The whole process has been repeatedly used to find a total of 158,719 blogs.

The main advantage of this methodology is that it takes benefit from the robust architecture of *wordpress.com*, a leading blog platform, as content- and language-filtering are outsourced, which seems to be efficient.

The discrepancy between the advertised and the actual number of blogs can be explained by the lack of incoming links or tags, to a substantial proportion of closed or restricted access blogs, and finally by the relative short crawl of *wordpress.com* with respect to politeness rules used.

2.2 Blog discovery in the wild

A detection phase is needed to be able to observe bloggers "in the wild" without needing to resort to large-scale crawling. In fact, guessing if a website uses WordPress by analysing HTML code is straightforward if nothing was been done to hide it, which is almost always the case. However, downloading even a reasonable number of web pages may take a lot of time. That is why other techniques have to be found to address this issue.

The detection process is twofold, the first filter is URL-based whereas the final selection uses HTTP HEAD requests. The permalinks settings defines five common URL structures for sites powered by WordPress, as well as a vocabulary to write customized ones. A HEAD request fetches the meta-information written in response headers without downloading the actual content, which makes it much faster, but also more resource-friendly, as less than three requests per domain name are sufficient.

Finally, the selection is made using a hard-coded decision tree, and the results are processed using the FLUX-toolchain, Filtering and Language identification for URL Crawling Seeds (Barbaresi, 2013a; Barbaresi, 2013b), which includes obvious spam and non-text documents filtering, redirection checks, collection of host- and markup-based data, HTML code stripping, document validity check, and language identification.

2.3 Content under CC-license

CC-licenses are increasingly popular public copyright licenses that enable the free distribution of an otherwise copyrighted work. A simple way to look for content under CC-licenses resides in scanning for links to the Creative Commons website, which proves to be relatively efficient, and is also used for instance by Lyding et al. (2014). We obtained similar results, with a very good recall and an precision around .65, with can be considered as being acceptable in this context.

That said, as a notable characteristic of internet content republishing resides in the severe copyright restrictions and potential penalties, we think that each and every blog that is scheduled for collection has to be carefully verified, an approach in which we differ from Lyding et al. (2014).

We describe the results of the manual evaluation phase in the evaluation section below. The results of automatic homepage scans on German blogs hosted by *wordpress.com* show that blogs including comments are rather rare, with 12,7% of the total (20,181 websites); 0,8% *at best* under CC license (1,201); and 0,2% *at best* with comments and under CC license (324).

To allow for blog discovery, large URL lists are needed. They were taken out previous web-crawling projects as well as out pages downloaded from *wordpress.com*. We obtained the following yields. There are more than 10e8 URLs URLs from the CommonCrawl project, of which approximately 1500 blogs mostly written in German and potentially under CC-license. The German Wikipedia links to more than 10e6 web documents outside of the Wikimedia websites, in which 300 potential targets were detected. In a list of links shared on social networks containing more than 10e3 different domain names, about 100 interesting ones were found. Last, there were

http://de.wordpress.com/tags/
Such as http://de.wordpress.com/tag/gesellschaft/
http://www.w3.org/Protocols/rfc2616/rfc2616
http://codex.wordpress.org/Using_Permalinks

http://creativecommons.org/licenses/
http://commoncrawl.org

more than 10e6 different URLs in the pages retrieved from *wordpress.com*, in which more than 500 potentially interesting blogs were detected.

In terms of yield, these results show that it is much more efficient to target a popular blog platform. Social networks monitoring is also a good option. Both yield understandably much more blog links than general URL lists. Even if large URL lists can compete with specific search with respect to the number of blogs discovered, they are much more costly to process. This finding consolidates the conclusions of Barbaresi (2014) concerning the relevance of the starting point of a crawl. In short, long crawls have a competitive edge as regards exhaustiveness, but it comes at a price.

The final list of blogs comprises 2727 candidates for license verification, of which 1218 are hosted on *wordpress.com* (45%).

3 Manual assessment of content and licenses

Blog classification has been performed manually using a series of predefined criteria dealing with (1) general classification, (2) content description, and (3) determination of authorship.

First, concerning the general classification, the essential criteria are whether there is really something to see on the page (e.g. no tests such as *lorem ipsum*) and whether it is really a blog. Another classification factor is whether the blog has been created or modified recently (i.e. after 2010-01-01).

Second, concerning the content description, the sine qua nons are to check that the page contains texts, a majority of which being in German, and that the text content is under a CC license. Other points are whether the webpage appears to be spam, whether the content can clearly be classified as dealing with Germany, Switzerland or Austria, whether the content appears to be *Hochdeutsch* or a particular dialect/sociolect, and last if the website targets a particular age group such as kids or young adults.

Third, the authorship criteria are twofold: is the blog a product of paid, professional editing or does it appear to be a hobby; and is the author clearly a woman, a man or a collective?

Concerning the essential criteria, the results of the classification are that 1,766 blogs can be used without restriction (65%), since all the textual content qualifies for archiving, meaning that there is text on the webpage, that it is a blog (it contains posts), that it is mostly written in German and that it is under CC license.

BY-NC-SA	652
BY-NC-ND	532
BY-SA	351
BY	282
BY-NC	129
BY-ND	58

Table 1: Most frequent license types

DE	1497
Unknown	715
AT	146
CH	69
LU	2
NL	2

Table 2: Most frequent countries (ISO code)

The breakdown of license types is shown in table 1, so are the results of country classification in table 2. The CC licensing can be considered to be a sure fact, since theoretically the CC license cannot be overridden once the content has been published. Possible differences between adaptations of the license in the various countries should not be an issue either, because it is done in a quite homogeneous way. The relatively high proportion of BY-NC-ND licenses (30%) is remarkable. While the "-ND" (no derivative works) restriction does not hinder republication as such, its compatibility with corpus building and annotation is unclear, so that such texts ought to be treated with caution.

4 Quantitative evaluation and comparison

4.1 Materials

We present a series of statistical analyses to get a glimpse of the characteristics of the crawled corpora. Content is divided into two different parts, the blog posts (BP), and the blog comments (BC), which do not necessarily share authorship. Due to

the relatively slow download of the whole blogs due to crawling politeness settings, we analyzed a subset of 696 blogs hosted on *wordpress.com* and 280 other WordPress blogs. We cannot calculate how synchronous the subtitles are with the blogs, manual analysis reveals a high proportion of TV series broadcast in the last few years.

Newspaper corpus

The results are compared with established text genres. On one hand, a newspaper corpus which is supposed to represent standard written German, extracted from the weekly newspaper *Die ZEIT*, more precisely the *ZEIT online* section (ZO), which features texts dedicated to online publishing. On the contrary, newspaper articles are easy to date, and we chose to use a subset ranging from 2010 to 2013 inclusive, which roughly matches both size and writing dates of the blogs. There have been digitally generated and are free of detection errors typical for retro-digitized newspaper corpora. ZO is in general considered to be a medium aiming at well-educated people. Therefore, we have picked it as a corpus representing standard educated German.

Subtitle corpus

On the other hand, a subtitle corpus (OS) which is believed to offer a more down-to-earth language sample. The subtitles were retrieved from the OpenSubtitles project, a community-based web platform for the distribution of movie and video game subtitles, then they were preprocessed and quality controlled (Barbaresi, 2014b). Subtitles as linguistic corpora have gained attention by the work of Brysbaert and colleagues (Brysbaert and New, 2009) who showed word frequencies extracted from movie subtitles were superior to frequencies from classical sources in explaining variance in the analysis of reaction times from lexical decision experiments. The reason for this superiority is still somewhat unclear (Brysbaert et al., 2011). It may stem from the fact that subtitles resemble spoken language, while traditional corpora are mainly compiled from written language (Heister and Kliegl, 2012). The analogy between subtitles and spoken language was also the primary motivation to include the OpenSubtitles corpus in the following analyses.

The corpora used in this study are all corpora from the Web. Structural properties of the corpora are shown in table 3. Their sizes are roughly comparable.

4.2 Preprocessing and Annotation

All corpora have been automatically split into tokens and sentences with the help of WASTE, Word and Sentence Tokenization Estimator (Jurish and Würzner, 2013), a statistical tokenizing approach based on a Hidden Markov Model (HMM), using the standard DTiger model. Subsequently, the resulting tokens have been assigned with possible PoS tags and corresponding lemmas by the morphological analysis system TAGH (Geyken and Hanneforth, 2006). The HMM tagger *moot* (Jurish, 2003) has then selected the most probable PoS tag for each token given its sentential context. In cases of multiple lemmas per best tag we chose the one with the lowest edit distance to the original token's surface.

4.3 Analyses

All corpora are aggregated on the level of types, lemmas and annotated types (i.e. type-PoS-lemma triplets) resulting in three different frequency mappings per corpus. Analyses are carried out using the statistical computing environment **R** (R Core Team, 2012).

Quantitative Corpus Properties

Table 3 summarizes a number of standard corpus characteristics. Token and type counts as well as length measures include punctuation. While token length is comparable in all four corpora, sentences in the subtitles are less than half as long as in the other corpora. The proportion of unknown types with respect to the standard-oriented morphological analyzer TAGH is by far smaller in the ZEIT corpus and marginally higher in blog comments than in the other standard-deviating corpora.

Type-Token Ratio

Figure 1 shows the number of types in the four examined corpora as a function of the size of growing corpus samples.

The number of different words with in a corpus is usually interpreted as a measure of its lexical

http://opensubtitles.org

Corpus	Size	⌀ TL	⌀ SL	unkn. T
Token level				
BP	33.0	4.95	20.3	2.76
BC	12.8	4.68	16.0†	2.75
ZO	38.2	5.08	17.5	0.89
OS	67.2	3.90	7.6	1.31
Type level				
BP	1.10	11.3	n/a	24.4
BC	0.56	10.5	n/a	27.3
ZO	0.98	12.2	n/a	13.7
OS	0.83	10.1	n/a	23.9

Size ... Number of tokens (resp. types) in the corpus in millions
TL ... Length of token (resp. type) in characters
SL ... Length of sentences in tokens
unkn. T ... Proportion of tokens (resp. types) unknown to TAGH

† Sentence length was re-computed using a statistical tokenization model (Jurish and Würzner, 2013) trained on the Dortmund Chat Corpus (Beißwenger, 2007). The original value using the standard newspaper model was 22.5, a dubious value.

Table 3: Various properties of the examined corpora.

variance. The plot shows that the OpenSubtitles corpus has a much smaller vocabulary than the three other corpora which are clearly dominated by the blog posts in this respect.

PoS Distribution

Table 4 lists percentage distributions for selected PoS tags on the level of tokens and types. We aggregated some of PoS categories for practical reasons. The figures show that the corpora are rather close in terms of tag distribution with a few remarkable differences. The higher amounts of pronouns and verbs in the subtitles is a direct consequence of shorter sentences. While the proportion of common names drops accordingly, this is not the case for the proper nouns, which validates the hypothesis that the subtitles actually replicate characteristics of spoken language. Besides, the lower proportion of common nouns and higher proportion of proper nouns in the blog comments indicates that it is relevant to study vocabulary diversity.

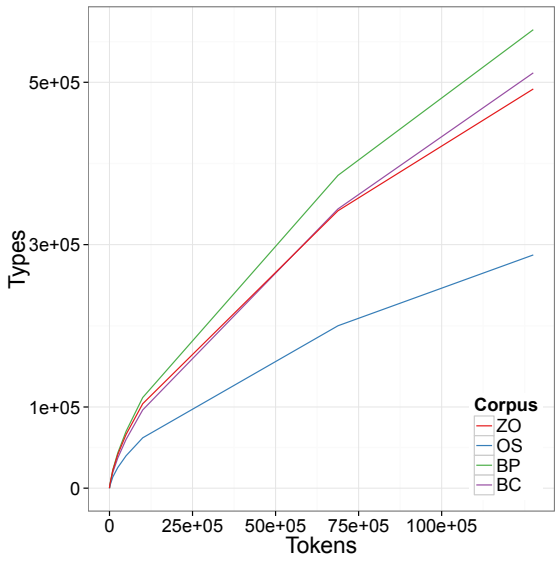

Figure 1: Number of types within random corpus samples (mean, 30 times iterated).

PoS \ Crps.	BP	BC	ZO	OS
Content words				
NN	16; 46	13; 42	18; 56	11; 42
NE	3; 22	2; 26	4; 18	3; 27
V*	12; 6	14; 8	13; 6	17; 9
AD*	14; 13	16; 14	13; 14	10; 11
Function words				
ART	8	6	10	5
AP*	8	7	8	4
P*	12	15	12	22
K*	5	5	4	3

Table 4: Percentage distribution of selected PoS (super)tags on token (content and function words) and type level (only content words). PoS tags are taken from the STTS. Aggregation of PoS categories is denoted by a wildcard asterisk. All percentages for function words on the type level are below one percent.

Frequency Correlations

For types shared by all evaluation corpora, Figure 2 shows correlations of their frequencies subdivided by frequency class. Frequency within the OpenSubtitles serves as the reference for frequency class since it is the largest corpus.

Correlations of subtitle frequencies with those from other corpora are clearly weaker than the other correlations while correlations of blog posts

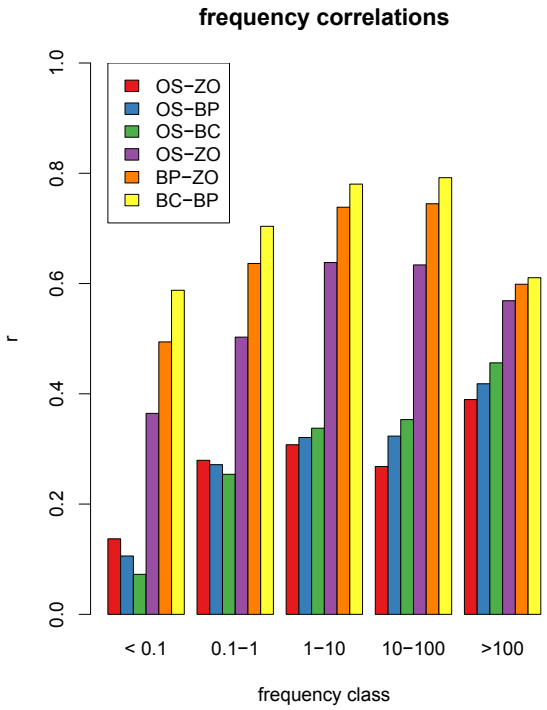

Figure 2: Correlations of type frequencies in different frequency classes.

and comments are always higher. The general pattern is the same in all frequency classes but the differences between the single correlation values are smaller in the highest and lowest range.

Vocabulary Overlap

Figure 3 shows overlaps in the vocabulary of the four corpora using a proportional Venn diagram (Venn, 1880). It has been generated using the *Vennerable* (Swinton, 2009) **R** package which features proportional Venn diagrams for up to nine sets using the Chow-Ruskey algorithm (Chow and Ruskey, 2004). The diagram is arranged into four levels each corresponding to the number of corpora sharing a type. The yellow layer contains types which are unique to a certain corpus. Types shared by two corpora are mapped to light orange levels while dark orange levels contain types shared by three corpora. Types present in all four corpora constitute the central red zone. The coloring of the borders of the planes denotes the involved corpora. In order to abstract from the different size of the data sets involved and to allow for an intuitive comparison of the proportions within the diagram, we included only the 100,000 most frequent words from each evaluation corpus into the analysis.

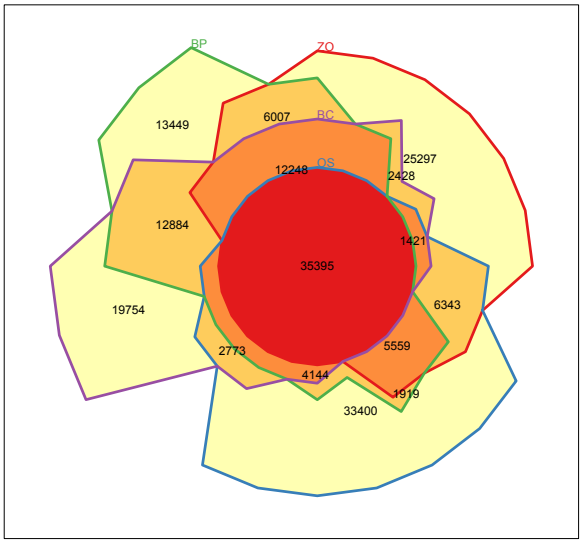

Figure 3: Venn diagram for the 100,000 most frequent words from each evaluation corpus.

Despite the heterogeneous nature of the corpora, there is a large overlap of roughly a third of the types between the four samples (red plane). Each sample contains a significant amount of exclusive tokens. The overlap between blog posts and comments is by far the largest on the second level while the one between blog posts and subtitles is the smallest. There is also a surprisingly large overlap between blog posts, comments and the ZEIT.

4.4 Discussion

The analyses above show large differences between the OpenSubtitles corpus on one and the ZEIT corpus on the other hand. These differences concern sentence length with much shorter sentences in the OS corpus; the amount of unknown words which includes non-standard word forms and (less frequent) named entities; frequency correlations which shows large frequency deviations in the medium frequency range and PoS distributions with fewer nouns and more verbs for the subtitles. We interpret these results as resembling some of the differences between spoken and written language.

In almost all analyses, blog content is found to be closer to the ZEIT corpus then to the OpenSubtitles corpus. This might be expected for the posts but it is somewhat surprising concerning the comments which are to a great extent discourse-like communication. Nonetheless, our quantitative results are in accordance with qualitative results on that matter (Storrer, 2001; Dürscheid, 2003).

In exception to that pattern, the amount of tokens unknown to TAGH in the blog samples is comparable to the value for the OpenSubtitles. This is caused by phenomena such as typos, standard-deviating orthography and *netslang* frequently observed in computer-mediated text and communication. In order to guarantee reliable linguistic annotation of blog posts and comments, emphasis will have to be put on improving existing and developing specific methods for automatic linguistic analysis.

5 Conclusion

First of all, our results show that it is possible to find blogs in German under Creative Commons license. The crawling and extraction tools seem to give a reasonable image of blog language, despite the fact that the CC license restriction impedes exploration in partly unknown ways and probably induces sociological biases.

We introduced evidence to try to classify blog corpora. Post content and comments seem to be different in nature, so that there is a real interest in separate analysis, all the more since it is possible to perform text extraction and linguistic annotation efficiently enough to allow for a comparison with more traditional or established text types. In this regard, a corpus comparison gives insights on distributional properties of the processed web texts.

Despite the presence of atypical word forms, tokens and annotation UFOs, most probably caused by language patterns typically found on the Internet, token-based analysis of blog posts and comments seems to bring these corpora closer to existing written language corpora.

More specifically, out-of-vocabulary tokens with respect to the morphological analysis are slightly more frequent in blog comments than in the other studied corpora. Concerning the lexical variance, blog posts dominate clearly, even if the higher proportion of proper nouns in the blog comments signalizes a promising richness regarding linguistic studies. Vocabulary overlap is best between blog posts and comments. However, a slight difference subsists between them, the latter being potentially closer to subtitles, as the PoS tag distribution seems to corroborate the hypothesis that subtitles are close to spoken language.

We believe that the visualizations presented in this article can help to answer everyday questions regarding corpus adjustments as well as more general research questions such as the delimitation of web genres.

Future work includes updates of the resources as well as full downloads of further blogs. Longer crawls as well as tries on other blog platforms might be a productive way to build bigger and potentially more diverse transmissible corpora. Additionally, more detailed annotation steps could allow for a thorough interpretation.

Part of the processing toolchain used in the experiments is available online under an open-source license. The corpora mentioned in this paper are available upon request.

Acknowledgments

Blog classification has been performed by Sophie Arana. Bryan Jurish has helped with the fine-tuning of the linguistic processing chain.

References

Yuki Arase and Ming Zhou. 2013. Machine Translation Detection from Monolingual Web-Text. In *Proceedings of the 51th Annual Meeting of the ACL*, pages 1597–1607.

Adrien Barbaresi. 2013a. Challenges in web corpus construction for low-resource languages in a post-BootCaT world. In Zygmunt Vetulani and Hans Uszkoreit, editors, *Proceedings of the 6th Language & Technology Conference, Less Resourced Languages special track*, pages 69–73.

Adrien Barbaresi. 2013b. Crawling microblogging services to gather language-classified URLs. Workflow and case study. In *Proceedings of the 51th Annual Meeting of the ACL, Student Research Workshop*, pages 9–15.

Adrien Barbaresi. 2014a. Finding Viable Seed URLs for Web Corpora: A Scouting Approach and Comparative Study of Available Sources. In Roland

https://github.com/adbar

Schäfer and Felix Bildhauer, editors, *Proceedings of the 9th Web as Corpus Workshop*, pages 1–8.

Adrien Barbaresi. 2014b. Language-classified Open Subtitles (LACLOS): download, extraction, and quality assessment. Technical report, BBAW. https://purl.org/corpus/german-subtitles.

Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Angelika Storrer. 2013. DeRiK: A German reference corpus of computer-mediated communication. *Literary and Linguistic Computing*, 28(4):531–537.

Michael Beißwenger. 2007. Corpora zur computervermittelten (internetbasierten) Kommunikation. *Zeitschrift für germanistische Linguistik*, 35(3):496–503.

Chris Biemann, Felix Bildhauer, Stefan Evert, Dirk Goldhahn, Uwe Quasthoff, Roland Schäfer, Johannes Simon, Leonard Swiezinski, and Torsten Zesch. 2013. Scalable Construction of High-Quality Web Corpora. *Journal for Language Technology and Computational Linguistics*, pages 23–59.

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M Jacobs, Jens Bölte, and Andrea Böhl. 2011. The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, 58(5):412–424.

Stirling Chow and Frank Ruskey. 2004. Drawing area-proportional venn and euler diagrams. In Giuseppe Liotta, editor, *Graph Drawing*, volume 2912 of *Lecture Notes in Computer Science*, pages 466–477. Springer.

Christa Dürscheid. 2003. Medienkommunikation im Kontinuum von Mündlichkeit und Schriftlichkeit. Theoretische und empirische Probleme. *Zeitschrift für Angewandte Linguistik*, 38:35–54.

Alexander Geyken and Thomas Hanneforth. 2006. TAGH: A Complete Morphology for German based on Weighted Finite State Automata. In *Finite State Methods and Natural Language Processing*, volume 4002 of *Lecture Notes in Computer Science*, pages 55–66. Springer.

Alexander Geyken. 2007. The DWDS corpus: A reference corpus for the German language of the 20th century. In Christiane Fellbaum, editor, *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*, pages 23–41. Continuum Press.

Julian Heister and Reinhold Kliegl. 2012. Comparing word frequencies from different German text corpora. In Kay-Michael Würzner and Edmund Pohl, editors, *Lexical Resources in Psycholinguistic Research*, pages 27–44. Potsdam Cognitive Science Series. vol.3.

Bryan Jurish and Kay-Michael Würzner. 2013. Word and Sentence Tokenization with Hidden Markov Models. *JLCL*, 28(2):61–83.

Bryan Jurish. 2003. A Hybrid Approach to Part-of-Speech Tagging. Final report, Kollokationen im Wörterbuch, Berlin-Brandenburgische Akademie der Wissenschaften.

Ben King and Steven Abney. 2013. Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods. In *Proceedings of NAACL-HLT*, pages 1110–1119.

R Core Team, 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org/.

Geoffrey Sampson. 2000. The role of taxonomy in language engineering. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 358(1769):1339–1355.

Roland Schäfer, Adrien Barbaresi, and Felix Bildhauer. 2013. The Good, the Bad, and the Hazy: Design Decisions in Web Corpus Construction. In Stefan Evert, Egon Stemle, and Paul Rayson, editors, *Proceedings of the 8th Web as Corpus Workshop*, pages 7–15.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt Cheap Web-Scale Parallel Text from the Common Crawl. In *Proceedings of the 51th Annual Meeting of the ACL*, pages 1374–1383.

Angelika Storrer. 2001. Getippte Gespräche oder dialogische Texte? Zur kommunikationstheoretischen Einordnung der Chat-Kommunikation. In Andrea Lehr, Matthias Kammerer, Klaus-Peter Konerding, Angelika Storrer, Caja Thimm, and Werner Wolski, editors, *Sprache im Alltag. Beiträge zu neuen Perspektiven in der Linguistik*, pages 439–466. De Gruyter.

Jonathan Swinton. 2009. Vennerable. http://r-forge.r-project.org/projects/vennerable.

John Venn. 1880. On the diagrammatic and mechanical representation of propositions and reasonings. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 10(59):1–18.