Challenges in the linguistic exploitation of specialized republishable web corpora

Adrien Barbaresi
Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)
barbaresi@bbaw.de

1 Context

Crawling and format I would like to present work on texts corpora in German, gathered on the Web and processed in order to be made available to linguists as well as to a broader user community via a web interface. The corpora are specialized in the sense that they only address a particular text genre or source at a time. As such, whether the texts are suitable or not for linguistic research is not an issue contrarily to general web corpora (Barbaresi, 2014), since the decision is made knowing what will be retrieved.

Web crawling techniques are used to download the documents, then they are stored in a processed version, i.e. not as web archives usually do (WARC format) but as a linguistic corpus in XML TEI format, in order to allow for a greater interoperability within the research community.

In this paper I describe two cases where texts are expected to be republishable: a "standard" case, German political speeches, and a "borderline" case, German blogs under Creative Commons license.

Institution and user base The work is performed in the context of the DWDS project, a digital dictionary of German (`dwds.de`). The DWDS has already gathered reference corpora of German texts from the 20th century and before (Geyken, 2007). The purpose of specific web corpora is to complement existing collections to allow for a better coverage of specific written text types and genres which are not found in traditional corpora, such as user-generated content as well as latest language evolutions.

Our primary user base consists of lexicographers, who need valuable or at least exploitable evidence, in the form of precise quotes or definition elements. There is a strong emphasis on metadata, since a quote without precise metadata (e.g. a date) is considered to be useless in the context of dictionary-making.

Construction and availability The actual gathering and processing of the corpora is described in previous publications (Barbaresi, 2012; Barbaresi & Würzner, 2014). Both are accessible online, under two different formats.

The political speeches have been made available both as archive and as browsable visualizations and documents.[1] The archive consists of files in XML TEI format which are meant to be used by specialists, while browser-friendly HTML documents provide a first glimpse of the corpus in terms of composition and actual content, most notably by a selection of relevant words and their chronological evolution throughout the corpus.

The blogs have been linguistically annotated with tools developed at the BBAW (Jurish & Würzner, 2013), they have been indexed and can be queried online [2], using bare words or more refined linguistic queries. A chronological view as well as visualizations of the corpus are under way. Additionally, it

[1] `http://purl.org/corpus/german-speeches`
[2] `http://kaskade.dwds.de/dstar/blogs/`

is planned to make the whole corpus available as downloadable archive under a CC license, since the licenses of its parts make it possible.

In the following, I focus on a series of challenges that are to be solved in order to make data from web archives accessible to researchers as well as to establish web text corpora as research objects: metadata extraction, quality assurance, licensing, and "scientificity".

2 Challenges

Challenge 1: Metadata extraction A proper metadata extraction is needed in order to make further downstream applications possible. It has to be performed meticulously, since experience shows that even small or rare mistakes in date encoding for instance may cause the application to be disregarded or discarded by researchers in the humanities, since linguistic trends cannot be identified properly if the content is not ordered in time.

However, content extraction is a real problem concerning large web corpora (Schäfer, Barbaresi, & Bildhauer, 2013), e.g. because of exotic markup and text genres, which is why potentially erroneous metadata in "one size fits all" web corpora may undermine the relevance of web texts for linguistic purposes.

In this particular case, easily available metadata in the case of speeches constrast with different content types, encodings, and markup patterns concerning the blogs. Compromises have to be made without sacrificing recall, since republishable texts are rather rare.

Challenge 2: Quality assessment of content Regarding the content, quality assurance is paramount, since a high quality is expected by users, all the more since they may feel reluctant to use web texts for their studies. In that sense, providing "Hi-Fi" web corpora also means promoting the cause of web sources and modernization of research methodology.

The quality has nothing to do with proper language or interesting content. It rather relies on formal criteria such as text integrity, and quality assessment of the cleaning and preprocessing steps, which can be performed semi-automatically by quantitative measures and specific visualizations (Barbaresi & Würzner, 2014).

Challenge 3: Licensing and republishing A simple way to look for content under Creative Commons licenses resides in scanning for URLs fragments involving CC licenses within a given web document, which proves to be relatively efficient (Lyding et al., 2014).

However, in order to ensure that the corpus is fully republishable under CC terms, a classification of blogs has been performed manually. Since the corpora are hosted in Germany, German copyright laws apply, which can be considered to be more restrictive than others, so that failure to exclude copyrighted text may be costly.

Content license is not a problem in the case of the political speeches, since they are considered to be public domain inasmuch as they have been read in public (this applies to Germany as well as to numerous countries).

Additionally, there are a number of issues with licensing in general and CC licenses in particular, even with manual verification: the "no derivative works" and (to a lesser extent) "non-commercial" predicates can hinder proper republication. The proportion of CC BY-NC-ND licenses in the blog corpus is relatively high (about 30%), which seems to exclude annotation and segmentation of the texts. There are also potential copyright issues regarding blog comments, whose status remains unclear.

Bottom line To sum up the issues described above, much work flows into ensuring the "scientificity" of web texts and making the texts not only available but also citable in a scholarly sense.

References

Barbaresi, A. (2012). *German Political Speeches – Corpus and Visualization* (Tech. Rep.). DGfS-CL Poster Session. (http://adrien.barbaresi.eu/corpora/speeches/)

Barbaresi, A. (2014). Finding Viable Seed URLs for Web Corpora: A Scouting Approach and Comparative Study of Available Sources. In R. Schäfer & F. Bildhauer (Eds.), *Proceedings of the 9th Web as Corpus Workshop* (pp. 1–8).

Barbaresi, A., & Würzner, K.-M. (2014). For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In *KONVENS 2014, NLP4CMC workshop proceedings* (pp. 2–10). Hildesheim University Press.

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In C. Fellbaum (Ed.), *Collocations and Idioms: Linguistic, lexicographic, and computational aspects* (pp. 23–41). Continuum Press.

Jurish, B., & Würzner, K.-M. (2013). Word and Sentence Tokenization with Hidden Markov Models. *JLCL*, *28*(2), 61–83.

Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell'Orletta, F., … Pirrelli, V. (2014). The paisa corpus of italian web texts. *Proceedings of the 9th Web as Corpus Workshop*, 36–43.

Schäfer, R., Barbaresi, A., & Bildhauer, F. (2013). The Good, the Bad, and the Hazy: Design Decisions in Web Corpus Construction. In S. Evert, E. Stemle, & P. Rayson (Eds.), *Proceedings of the 8th Web as Corpus Workshop* (pp. 7–15).