Two comparable corpora of German newspaper text gathered on the web: Bild & Die Zeit

Adrien Barbaresi

Two comparable corpora of German newspaper text gathered on the web: Bild & Die Zeit

Technical report

Adrien Barbaresi
ICAR lab / ENS Lyon

Contents

1 Summary

This technical report documents the creation of two comparable corpora of German newspaper text, focused on the daily tabloid *Bild* and the weekly newspaper *Die Zeit*.

Two specialized crawlers and corpus builders were designed in order to crawl the domain names bild.de and zeit.de with the objective of gathering as many complete articles as possible. A high content quality was made possible by the specially designed boilerplate removal and metadata recording code.

As a result, two separate corpora were created. Currently, the last version for *Bild* is from 2011 and the last version for *Die Zeit* is from early 2013. The corpora feature a total of respectively 60 476 and 134 222 articles.

Whereas the crawler designed for *Bild* has been discontinued due to frequent layout changes on the website, the other one concerning *Die Zeit* is still actively maintained, its code has been made available under an open source license.

2 Design decisions

2.1 Choice of the newspapers

The primary goal was the constitution of a reference corpus of comparable newspaper text. Both newspapers were chosen accordingly, since they accounted for two specific strategies in the general press.

At a glance there are striking differences in the writing style of the articles. In *Bild* there is a tendency towards very short sentences and the use of colons and illustrated content (multimedia components were not part of the crawling strategy). Comparatively, there is a lot more textual content in *Die Zeit* with a proclivity towards subordination and elaborate sentences as well as long and detailed articles. In short, it seems that *Bild* uses a lot more parataxis whereas there is a lot more hypotaxis to be found in *Die Zeit*.

The expected audiences of these newspapers are also quite different. While both can be considered as successful (with *Bild* being one of the most popular newspapers in Europe, selling about five times more copies than *Die Zeit*), the first one aims at keeping its leading status of popular tabloid, while the latter is well-regarded for its journalistic quality.

Last, there are also technical reasons explaining this choice: on both sides, crawling was and is not explicitly forbidden, which is not always the case by German newspaper websites. As a matter of fact, no hindrance was encountered while performing the crawls.

2.2 Crawling of the websites

Starting from the front page or from a given list of links, the crawlers retrieve newspaper articles and gather new links to explore them as it goes. The HTML code as well as the superfluous text are stripped in order to spare disk space, the remaining text (which is possibly the exact content of the article) is saved as a raw text file with relevant metadata (such as title, subtitle, excerpt, author, date and URL).

They also detect and filter out undesirable content based on a URL analysis, as the URLs give precious hints about the article column or about website specific and internal documents. In fact, a few columns were discarded because most of the texts they contained were of a different nature or genre, such as the cooking recipes of *Die Zeit* or the topless women pages on the website of *Bild*.

As the crawling process took place at a time when online versions of the newspapers emerged, it was impacted by the editorial changes related to this evolution and the duplication of the articles (see below) as well as the existence of near duplicate documents are symptomatic for the erratic decisions taken by the editorial staff, leading for example to the parallel existence of online and print versions on the website of *Die Zeit* in 2008.

2.3 Technicalities

Precision & recall Precision was preferred to recall, which means that getting complete articles was considered more important than gathering all articles without distinction.

Deduplication A hash function was used to shorten the links and make sure a given URL was retrieved just once. Nonetheless, there were still duplicate content for various reasons, including lack of overview on the publisher side and refactoring of articles under different titles. Titles, excerpts (if the case applied) and text hashes altogether were used to deduplicate the texts in the corpus. Further analysis may show that there are very similar texts in the corpus, which may be explained by genre-specific features.

Boilerplate removal The uninteresting (boilerplate) parts of a web page were cut off using specially crafted regular expressions.

Setting The crawlers are relatively fast (even if they were not set up for speed) and do not need a lot of computational resources. They may be run on a personal computer.

3 Current versions of the corpora

3.1 Bild corpus

There are 60 476 unique documents in the corpus, which were published between November 2007 and November 2010 according to the metadata. The corpus comprises 19 404 260 tokens.

3.2 Die Zeit corpus

There are 134 222 unique documents in the corpus, which were published between 1946 and January 2013 according to the metadata, with the major part of the articles being posterior to 2005. The corpus comprises 105 456 462 tokens.

4 References

4.1 Access to the corpora

So far the corpora have been used internally at the ENS Lyon. Although they cannot be republished as is, parts of it could be made available upon request, as well as derivative works such as n-gram lists.

4.2 Software

The pieces of code needed to crawl *Die Zeit* are available under an open source licence: https://code.google.com/p/zeitcrawler/
This includes a list of links as well as scripts to convert raw data into XML format for further use with natural language processing tools.