Efficient construction of metadata-enhanced web corpora

Adrien Barbaresi

Efficient construction of metadata-enhanced web corpora

Adrien Barbaresi
Austrian Academy of Sciences – Berlin-Brandenburg Academy of Sciences
adrien.barbaresi@oeaw.ac.at

Abstract

Metadata extraction is known to be a problem in general-purpose Web corpora, and so is extensive crawling with little yield. The contributions of this paper are threefold: a method to find and download large numbers of WordPress pages; a targeted extraction of content featuring much needed metadata; and an analysis of the documents in the corpus with insights of actual blog uses.

The study focuses on a publishing software (WordPress), which allows for reliable extraction of structural elements such as metadata, posts, and comments. The download of about 9 million documents in the course of two experiments leads after processing to 2.7 billion tokens with usable metadata. This comparatively high yield is a step towards more efficiency with respect to machine power and "Hi-Fi" web corpora.

The resulting corpus complies with formal requirements on metadata-enhanced corpora and on weblogs considered as a series of dated entries. However, existing typologies on Web texts have to be revised in the light of this hybrid genre.

1 Introduction

1.1 Context

This article introduces work on focused web corpus construction with linguistic research in mind. The purpose of focused web corpora is to complement existing collections, as they allow for better coverage of specific written text types and genres, user-generated content, as well as latest language evolutions. However, it is quite rare to find ready-made resources. Specific issues include first the discovery of relevant web documents, and second the extraction of text and metadata, e.g. because of exotic markup and text genres (Schäfer et al., 2013). Nonetheless, proper extraction is necessary for the corpora to be established as scientific objects, as science needs an agreed scheme for identifying and registering research data (Sampson, 2000). Web corpus yield is another recurrent problem (Suchomel and Pomikálek, 2012; Schäfer et al., 2014). The shift from web *as* corpus to web *for* corpus – mostly due to an expanding Web universe and the need for better text quality (Versley and Panchenko, 2012) – as well as the limited resources of research institutions make extensive downloads costly and prompt for handy solutions (Barbaresi, 2015).

The DWDS lexicography project[1] at the Berlin-Brandenburg Academy of Sciences already features a good coverage of specific written text genres (Geyken, 2007). Further experiments including internet-based text genres are currently conducted in joint work with the Austrian Academy of Sciences (Academy Corpora). The common absence of metadata known to the philological tradition such as authorship and publication date accounts for a certain defiance regarding Web resources, as linguistic evidence cannot be cited or identified properly in the sense of the tradition. Thus, missing or erroneous metadata in "one size fits all" web corpora may undermine the relevance of web texts for linguistic purposes and in the humanities in general. Additionally, nearly all existing text extraction and classification techniques have been developed in the field of information retrieval, that is not with linguistic objectives in mind.

[1] Digital Dictionary of German, http://zwei.dwds.de

The contributions of this paper are threefold:
(1) a method to find and download large amounts of WordPress pages;
(2) a targeted extraction of content featuring much needed metadata;
(3) an analysis of the documents in the corpus with insights of actual uses of the blog genre.
My study focuses on a publishing software with two experiments, first on the official platform wordpress.com and second on the .at-domain. WordPress is used by about a quarter of the websites worldwide[2], the software has become so broadly used that its current deployments can be expected to differ from the original ones. A number of 158,719 blogs in German have previously been found on wordpress.com (Barbaresi and Würzner, 2014). The .at-domain (Austria) is in quantitative terms the 32th top-level domain with about 3,7 million hosts reported.[3]

1.2 Definitional and typological criteria

From the beginning of research on blogs/weblogs, the main definitional criterion has always been their form, a "reverse chronological sequences of dated entries" (Kumar et al., 2003). Another formal criterion is the use of dedicated software to articulate and publish the entries, a "weblog publishing software tool" (Glance et al., 2004), "public-domain blog software" (Kumar et al., 2003), or Content Management System (CMS). These tools largely impact the way blogs are created and run. 1996 seems to be the acknowledged beginning of the blog/weblog genre, with an exponential increase of their use starting in 1999 with the emergence of several user-friendly publishing tools (Kumar et al., 2003; Herring et al., 2004).

Whether a blog is to be considered to be a web page in its whole (Glance et al., 2004) or a website containing a series of dated entries, or posts, (Kehoe and Gee, 2012) being each a web page, there are invariant elements, such as "a persistent sidebar containing profile information" as well as links to other blogs (Kumar et al., 2003), or blogroll. For that matter, blogs are intricately intertwined in what has been called the blogosphere: "The cross-linking that takes place between blogs, through blogrolls, explicit linking, trackbacks, and referrals has helped create a strong sense of community in the weblogging world." (Glance et al., 2004). This means that a comprehensive crawl could lead to better yields.

Regarding the classification of blogs, Blood (2002) distinguishes three basic types: filters, personal journals, and notebooks, while Krishnamurthy (2002) builds a typology based on function and intention of the blogs: online diaries, support group, enhanced column, collaborative content creation. More comprehensive typologies established on one hand several genres: online journal, self-declared expert, news filter, writer/artist, spam/advertisement; and on the other hand distinctive "variations": collaborative writing, comments from readers, means of publishing (Glance et al., 2004).

2 Related work

2.1 (Meta-)Data Extraction

Data extraction has first been based on "wrappers" (nowadays: "scrapers") which were mostly relying on manual design and tended to be brittle and hard to maintain (Crescenzi et al., 2001). These extraction procedures have also been used early on by blogs search engines (Glance et al., 2004). Since the genre of "web diaries" was established before the blogs in Japan, there have been attempts to target not only blog software but also regular pages (Nanno et al., 2004), in which the extraction of metadata also allows for a distinction based on heuristics.

Efforts were made to generate wrappers automatically, with emphasis on three different approaches (Guo et al., 2010): wrapper induction (e.g. by building a grammar to parse a web page), sequence labeling (e.g. labeled examples or a schema of data in the page), and statistical analysis and series of resulting heuristics. This analysis combined to the inspection of DOM tree characteristics (Wang et al., 2009; Guo et al., 2010) is a common ground to the information retrieval and web corpus linguistics communities, with the categorization of HTML elements and linguistic features (Ziegler and Skubacz, 2007) for the former, and markup and boilerplate removal operations known to the latter community (Schäfer and Bildhauer, 2013).

Regarding content-based wrappers for blogs in particular, targets include the title of the entry, the date, the author, the content, the number of comments, the archived link, and the trackback link (Glance et al., 2004); they can also aim at com-

[2] http://w3techs.com/technologies/details/cm-wordpress/all/all

[3] http://ftp.isc.org/www/survey/reports/2016/01/bynum.txt

ments specifically (Mishne and Glance, 2006).

2.2 Blog corpus construction

The first and foremost issue in blog corpus construction still holds true today: "there is no comprehensive directory of weblogs, although several small directories exist" (Glance et al., 2004). Previous work established several modes of construction, from broad, opportunistic approaches, to the focus on a particular method or platform due to the convenience of retrieval processes. Corpus size and length of downloads are frequently mentioned as potential obstacles. Glance et al. (2004) performed URL harvesting through specialized directories, and found a practical upper bound at about 100,000 active weblogs, which were used as a corpus in their study.

The first comprehensive studies used feeds to collect blog texts (Gruhl et al., 2004), since they are a convenient way to bypass extensive crawling and to harvest blog posts (and more rarely comments) without needing any boilerplate removal.

An approach based on RSS and Atom feeds is featured in the TREC-Blog collection[4] (Macdonald and Ounis, 2006), a reference in Information Extraction which has been used in a number of evaluation tasks. 100,649 blogs were predetermined, they are top blogs in terms of popularity, but no further information is given. Spam blogs, and hand-picked relevant blogs (no information on the criteria either) are used to complement and to balance the corpus to make it more versatile. The corpus is built by fetching feeds describing recent postings, whose permalinks are used as a reference. From initial figures totaling 3,215,171 permalinks and 324,880 homepages, most recent ones from 2008 mention 1,303,520 feeds and 28,488,766 permalink documents.[5]

Another way to enhance the quality of data and the ease of retrieval is the focus on a particular platform. To study authorship attribution, Schler et al. (2006) gathered a total of 71,000 blogs on the Google-owned Blogger platform, which allowed for easier extraction of content, although no comments are included in the corpus.

The Birmingham Blog Corpus (Kehoe and Gee, 2012) is a more recent approach to comprehensive corpus construction. Two platforms are taken into consideration: Blogger and wordpress.com, with the "freshly pressed" page on WordPress as well as a series of trending blogs used as seed for the crawls, leading to 222,245 blog posts and 2,253,855 comments from Blogger and WordPress combined, totaling about 95 million tokens (for the posts) and 86 million tokens (for the comments).

The YACIS Corpus (Ptaszynski et al., 2012) is a Japanese corpus consisting of blogs collected from a single blog platform, which features mostly users in the target language as well as a clear HTML structure. Its creators were able to gather about 13 million webpages from 60,000 bloggers for a total of 5.6 billion tokens.

Last, focused crawl on the German version of the platform wordpress.com led to the construction of a corpus of 100 million tokens under Creative Commons licenses (Barbaresi and Würzner, 2014), albeit with a much lower proportion of comments (present on 12.7% of the posts). In fact, comments have been shown to be strongly related to the popularity of a blog (Mishne and Glance, 2006), so that the number of comments is much lower when blogs are taken at random.

The sharp decrease in publication of work documenting blog corpus construction after 2008 signalizes a shift of focus, not only because web corpus construction does not often get the attention it deserves, but also because of the growing popularity of short message services like Twitter, which allow for comprehensive studies on social networks and internet-based communication, with a larger number of users and messages as well as clear data on network range (e.g. followers).

3 Method

3.1 Discovery

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 analyzing 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 I chose to perform massive scans in order to find websites using WordPress, which to my best knowledge has not yet been tried in the literature. The detection process is twofold, the first filter is URL-based whereas the final selection uses shallow HTTP requests.

[4] http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG

[5] https://web.archive.org/web/20160313020503/http://ir.dcs.gla.ac.uk/test_collections/blogs08info.html

The permalinks settings[6] in WordPress define five common URL structures: *default* (?p= or ?page_id= or ?paged=), *date* (/year/ and/or /month/ and/or /day/ and so on), *post number* (/keyword/number – where keyword is for example "archives"), *tag or category* (/tag/, /category/, or cross-language equivalents), and finally *post name* (long URLs containing a lot of hyphens). Patterns derived from those structures can serve as a first filter, although the patterns are not always reliable: news websites tend to use dates very frequently in URLs, in that case the accuracy of the prediction is poor.

The most accurate method would be a scan of fully-rendered HTML documents with clear heuristics such as the "generator" meta tag in the header, which by default points to WordPress. In this study, HTTP HEAD[7] requests are used to spare bandwidth and get cleaner, faster results. HEAD requests are part of the HTTP protocol. Like the most frequent request, GET, which fetches the content, they are supposed to be implemented by every web server. 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 according to my method less than three requests per domain name are sufficient.

The following rules come from from the official documentation and have been field-tested:
(1) A request sent to the homepage is bound to yield pingback information to use via the XML-RPC protocol in the *X-Pingback* header. Note that if there is a redirect, this header usually points to the "real" domain name and/or path, ending in *xmlrpc.php*. What is more, frequently used WordPress modules may leave a trace in the header as well, e.g. *WP-Super-Cache*, which identifies a WordPress-run website with certainty.
(2) A request sent to */login* or */wp-login.php* should yield a HTTP status corresponding to an existing page (2XX, 3XX, more rarely 401).
(3) A request sent to */feed* or */wp-feed.php* should yield the header *Location*.

The criteria can be used separately or in combination. I chose to use a simple decision tree. The information provided is rarely tampered on or misleading, since almost all WordPress installations stick to the defaults. Sending more than one request makes the guess more precise, it also acts like a redirection check which provides the effectively used domain name behind a URL. Thus, since the requests help deduplicating a URL list, they are doubly valuable.

3.2 Sources and crawls

This study falls doubly into the category of focused or scoped crawling (Olston and Najork, 2010): the emphasis lies on German or on the .at-domain, and a certain type of websites are examined based on structural characteristics.

I have previously shown that the diversity of sources has a positive impact on yield and quality (Barbaresi, 2014). Aside from URL lists from this and other previous experiments (Barbaresi, 2013) and URLs extracted from each batch of downloaded web documents (proper crawls), several sources were queried, not in the orthodox BootCat way with randomized tuples (Baroni and Bernardini, 2004) but based on formal URL characteristics as described above:
(1) URLs from the CommonCrawl[8], a repository already used in web corpus construction (Habernal et al., 2016; Schäfer, 2016);
(2) the CDX index query frontend of the internet Archive;[9]
(3) public instances of the metasearch engine Searx.[10]

A further restriction resides in the downloads of sitemaps for document retrieval. A majority of websites are optimized in this respect, and experiments showed that crawls otherwise depend on unclear directory structures such as posts classified by categories or month, as well as on variables (e.g. *page*) in URL structures, which leads to numerous duplicates and an inefficient crawl. Another advantage is that websites offering sitemaps are almost systematically robot-friendly, which solves ethical robots.txt-related issues such as the crawl delay, which is frequently mentioned as an obstacle in the literature.

3.3 Extraction

I designed a text extraction targeting specifically WordPress pages, which is transferable to a whole range of self-hosted websites using WordPress, allowing to reach various blogger profiles thanks

[6] http://codex.wordpress.org/Using_Permalinks
[7] http://www.w3.org/Protocols/rfc2616/rfc2616
[8] http://commoncrawl.org
[9] https://github.com/internetarchive/wayback/tree/master/wayback-cdx-server
[10] https://github.com/asciimoo/searx/wiki/Searx-instances

to a comparable if not identical content structure. The extractor acts like a state-of-the-art wrapper: after parsing the HTML page, XPATH-expressions select subtrees and operate on them through pruning and tag conversion to (1) write the data with the desired amount of markup and (2) convert the desired HTML tags into the output XML format in strict compliance to the guidelines of the Text Encoding Initiative[11], in order to allow for a greater interoperability within the research community.

The extraction of metadata targets the following fields, if available: title of post, title of blog, date of publication, canonical URL, author, categories, and tags. The multiple plugins cause strong divergences in the rendered HTML code, additionally not all websites use all the fields at their disposal. Thus, titles and canonical URL are the most often extracted data, followed by date, categories, tags, and author.

Content extraction allows for a distinction between post and comments, the latter being listed as a series of paragraphs with text formatting. The main difference with extractors used in information retrieval is that structural boundaries are kept (titles, paragraphs), whereas links are discarded for corpus use. A special attention is given to dates. Documents with non-existent or missed date or entry content are discarded during processing and are not part of the corpus, which through the dated entries is a corpus of "blogs" in a formal sense. Removal of duplicates is performed on entry basis.

3.4 Content analysis

In the first experiment, language detection is performed with langid.py (Lui and Baldwin, 2012) and sources are evaluated using the Filtering and Language identification for URL Crawling Seeds[12] toolchain (Barbaresi, 2014), 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. No language detection is undertaken is the second experiment since no such filtering is intended. That being said, a large majority of webpages are expected to be in German, as has been shown for another German-speaking country in the .de-TLD (Schäfer et al., 2013).

The token counts below are produced by the WASTE tokenizer (Jurish and Würzner, 2013).

[11] http://www.tei-c.org/

[12] https://github.com/adbar/flux-toolchain

4 Experiment 1: Retrieving German blogs

4.1 General figures on harvested content

In a previous experiment, the largest platform for WordPress-hosted websites, wordpress.com, blogs under CC license were targeted (Barbaresi and Würzner, 2014). In the summer of 2015, sitemaps were retrieved for all known home pages, which lead to the integral download of 145,507 different websites for a total number of 6,605,078 documents (390 Gb), leaving 6,095,630 files after processing (36 Gb). There are 6,024,187 "valid" files (with usable date and content) from 141,648 websites, whose text amounts to about 2.11 billion tokens.

The distribution of harvested documents in the course of years is documented in table 6, there are 6,095,206 documents with at least a reliable indication of publication year, i.e. 92.3% of all documents. Contrarily to dates in the literature, these results are not from reported permalinks dates from feeds, but directly from page metadata; nonetheless, there is also a fair share of implausible dates, comparable to the 3% of the TREC blog corpus (Macdonald and Ounis, 2006). This indicates that these dates are not an extraction problem but rather a creative license on the side of the authors.

Year	Docs.
2003	1,746
2004	4,993
2005	13,916
2006	62,901
2007	191,898
2008	377,271
2009	575,923
2010	733,397
2011	871,108
2012	1,066,996
2013	1,108,495
2014	717,861
2015	362,633
rest	6,068

Table 1: Distribution of documents among plausible years in the first experiment

4.2 Typology

An analysis of the top domain names in canonical URLs extracted from the documents, by total

number of documents in the corpus (see table 2) yields a typology clearly oriented towards community blogging mostly centered on creative activities or hobbies.

Domain name	URLs
mariusebertsblog.com	2,954
allesnaehbar.de	1,730
zuckerzimtundliebe.de	1,194
lyrikzeitung.com	1,104
der-retrosalon.com	1,092
rhein-main-startups.com	1,046
sciencefiles.org	1,042
des-belles-choses.com	1,014
twinklinstar.com	1,013
wirsindeins.org	1,007

Table 2: Most frequent domains names in URL queue in the first experiment

There are 4,777,546 pages with categories, 10,288,861 uses and 312,055 different categories in total, the top-15 results are displayed in table 3.

	Name	Freq.	Translation
1	Uncategorized	588,638	*(defaut category)*
2	Allgemein	239,796	*general*
3	Politik	71,534	*politics*
4	Allgemeines	60,178	*general*
5	News	58,281	*(also German)*
6	Musik	46,238	*music*
7	Gesellschaft	35,675	*society*
8	Fotografie	35,042	*photography*
9	Deutschland	33,841	*Germany*
10	Aktuelles	33,117	*current topics*
11	Medien	30,914	*media*
12	Alltag	29,839	*everyday life*
13	Leben	27,897	*life*
14	Fotos	26,107	*pictures*
15	Sonstiges	24,431	*misc.*

Table 3: Most frequent categories in the first experiment

There are 2,312,843 pages with tags, 15,856,481 uses in total, and 2,431,920 different tags, the top-15 results are displayed in table 4. They are as general as the top categories but slightly more informative.

All in all, the observed metadata are in line with the expectations, even if the high proportion of photoblogs is not ideal for text collection. Comments were extracted for 1,454,752 files (24%), this proportion confirms the hypothesis that the wordpress.com-platform leads primarily to the publication of blogs in a traditional fashion. On the contrary, the typology has to be more detailed in the second experiment due to the absence of previous knowledge about the collection.

	Name	Freq.	Translation
1	Fotografie	35,910	*photography*
2	Berlin	34,553	
3	Deutschland	30,351	*Germany*
4	Leben	29,597	*life*
5	Politik	26,315	*politics*
6	Musik	26,221	*music*
7	Foto	26,202	
8	Liebe	24,865	*love*
9	Kunst	24,382	*art*
10	USA	21,059	
11	Fotos	20,829	*pictures*
12	Natur	17,490	*nature*
13	Gedanken	16,542	*thoughts*
14	Weihnachten	16,344	*christmas*
15	Video	16,329	

Table 4: Most frequent tags in the first experiment

5 Experiment 2: Targeting the .at-domain

5.1 General figures on harvested content

The iterations summed up in table 5 took place during the 2nd half of 2015. Each time, all links in the .at top level domain were extracted, and analyzed as to their potential to be using WordPress. If so, potential sitemaps were retrieved and the URLs added to the queue if they were new. When necessary (e.g. after stages 5 and 6), the crawls have been refreshed with new material described in sources. After 11 iterations, seed exhaustion was nearing as new WordPress websites with sitemaps were hard to come by, and the experiment was stopped.

batch	domains	no. files	Gb
1	2,020	571,888	31
2	525	103,211	5.5
3	1,269	695,827	34
4	109	49,488	3.3
5	84	433	0.02
6	206	37,632	1.7
7	1,405	483,566	21
8	1,603	175,456	11
9	458	62,103	4.1
10	1,887	456,419	27
11	2,988	417,951	20

Table 5: Iterations and yields in the second experiment

A total of 3,053,974 different URLs have been downloaded (159 Gb), which left after processing and canonicalization 2,589,674 files (14 Gb). There are about 2 million "valid" files (with usable date and content), whose text amounts to about 550 million tokens. There are 5,664 different domain names before processing, and 7,275 after (due to the resolution of canonical URLs).

5.2 Typology

Of all canonical domain names, only 240 contain the word *blog*. Comments were extracted for 181,246 files (7%), which is explained mainly by the actual absence of comments and partly by difficulty of extraction in the case of third-party comment systems.

The distribution of harvested documents in the course of years is documented in table 6. There are 2,083,535 documents with at least a reliable indication of publication year, i.e. 80.5% of all documents. The relative amount of "creative" dates is slightly higher than in experiment 1, which hints at a larger diversity of form and content.

The increase in the number of documents exceeds by far the increase of domains registered in the .at-TLD[13], which seems to hint at the growing popularity of WordPress and maybe also at the ephemeral character of blogs.

Year	Docs.
2003	17,263
2004	30,009
2005	28,177
2006	35,853
2007	47,934
2008	78,895
2009	104,604
2010	152,422
2011	176,231
2012	197,819
2013	297,143
2014	371,605
2015	517,073
rest	28,507

Table 6: Distribution of documents among plausible years in the second experiment

An analysis of the top-50 domains names in canonical URLs extracted from the documents, by total number of documents in the corpus (see table 7) gives the following typology: informational for general news websites (9), promotional/commercial for websites which list ads, deals, jobs or products (12), specialized for focused news and community websites (16), entertainment (3), political (3), personal for websites dedicated to a person or an organization (3), adult (2), forum (1).

There are 260,468 pages with categories, 834,284 uses and 11,813 different categories in total, the top-15 results are displayed in table 8. The fact that "blog" is used as a category shows that it is not taken for granted.

There are 279,083 pages with tags, 5,093,088 uses in total, and 192,352 different tags, the top-15 results are displayed in table 9. The tags reflect a number of different preoccupations, including family, holidays, sex, job and labor legislation. "Homemade" and "amateur" can be used in German, albeit rarely, these words give more insights on the genre (most probably adult entertainment)

[13] 1,594,059 in January 2005; 3,112,683 in January 2010; 3,630,078 in January 2015
Source: http://ftp.isc.org/www/survey/reports/

Domain name	URLs	Genre
vol.at	333,690	informational
triple-s.at	312,714	informational
salzburg24.at	134,230	informational
vienna.at	96,654	informational
vorarlbergernachrichten.at	49,816	informational
dealdoktor.de	25,445	promotional
sportreport.biz	24,796	informational
cba.fro.at	21,895	informational
juve-verlag.at	21,548	promotional
eventfotos24.at	17,805	entertainment
unibrennt.at	16,497	political
dolomitenstadt.at	16,484	informational
sparhamster.at	13,997	promotional
freizeitalpin.com	12,440	specialized
webdeals.at	11,472	promotional
hans-wurst.net	10,717	entertainment
autorevue.at	9,840	specialized
katja.at	9,833	political
spielweb.at	9,541	promotional
medmedia.at	9,125	specialized
adiweiss.at	8,741	personal
sciam-online.at	8,058	specialized
electronicbeats.net	7,255	specialized
antiquariat-getau.at	7,205	promotional
greenpeace.org	7,031	political
photoboerse.at	6,945	promotional
salzburgresearch.at	6,802	professional
mittelstand-nachrichten.at	6,694	informational
sturm12.at	6,672	specialized
raketa.at	6,170	entertainment
platzpirsch.at	6,127	promotional
sexyinsider.at	6,024	adult
rebell.at	5,782	specialized
jusportal.at	5,739	specialized
aktuelle-veranstaltungen.at	5,733	specialized
ffmoedling.at	5,633	personal
zddk.eu	5,386	forum
kosmetik-transparent.at	5,381	specialized
sportwetteninfo.at	5,366	specialized
autoguru.at	5,142	specialized
ps4news.at	5,102	specialized
gastronomiejobs.wien	5,035	promotional
psychohelp.at	4,836	promotional
porno-austria.at	4,822	adult
christianmari.at	4,709	promotional
blog.sprachreisen.at	4,493	personal
w6-tabletop.at	4,488	specialized
ellert.at	4,381	promotional
demonic-nights.at	4,353	specialized
todesanzeigen.vol.at	4,296	specialized

Table 7: Most frequent domains names in URL queue in the second experiment

	Name	Freq.	Translation
1	Allgemein	28,005	*general*
2	Blu-ray	10,445	*(laser disc standard)*
3	MedienFamilie	9,662	*media-family*
4	Blog	9,652	
5	Familienleben	9,278	*family life*
6	News	8,857	*(also German)*
7	Film	8,222	*movies*
8	Absolut-Reisen	6,964	*absolute travels*
9	Buch	6,146	*book*
10	Schule	6,108	*school*
11	Spiele	5,939	*games*
12	Familienpolitik	5,781	*family policies*
13	Gewinnspiel	5,607	*competition*
14	In eigener Sache	5,463	*in our own cause*
15	Uncategorized	5,150	*(defaut category)*

Table 8: Most frequent categories in the second experiment

	Name	Freq.	Translation
1	Wien	18,973	*Vienna*
2	Deutschland	18,895	*Germany*
3	Usermeldungen	14,409	*user reports*
4	sterreich	10,886	*Austria*
5	Angebot aus DE	10,155	*offer from Germany*
6	sex	10,112	
7	Frauen	9,541	*women*
8	Kinder	8,968	*children*
9	USA	8,013	
10	Urlaub	7,767	*holiday*
11	homemade	7,666	
12	amateur	7,660	
13	mydirtyhobby	7,635	
14	Recht	7,611	*law*
15	Arbeitsrecht	7,294	*labor legislation*

Table 9: Most frequent tags in the second experiment

than on content language.

All in all, the distribution of categories and tags indicates that the majority of texts target as expected German-speaking users.

6 Discussion

Although the definition of blogs as a hybrid genre neither fundamentally new nor unique (Herring et al., 2004) holds true, several assumptions about weblogs cannot be considered to be accurate anymore in the light of frequencies in the corpus. Blogs are not always "authored by a single individual" (Kumar et al., 2003), nor does the frequency criterion given by the Oxford English Dictionary (Kehoe and Gee, 2012) – "frequently updated web site" – necessarily correspond to the reality. Even if both experiments gathered blogs in a formal sense, there are differences between the websites on the platform wordpress.com and freely hosted websites. The former are cleaner in form and content, they are in line with a certain tradition. The "local community interactions between a small number of bloggers" (Kumar et al., 2003) of the beginnings have been relegated by websites corresponding to the original criteria of a blog but whose finality is to sell information, entertainment, or concrete products and services.

Consequently, the expectation that "blog software makes Web pages truly interactive, even if that interactive potential has yet to be fully exploited" (Herring et al., 2004) is either outdated or yet to come. Beside these transformations and the emergence of other social networks, the whole range from top to barely known websites shows that the number of comments per post and per website is largely inferior to the "bursting" phase of webblogging, where comments were "a substantial part of the blogosphere" (Mishne and Glance, 2006). The evolution of the Web as well as the scope of this study cast the typical profile of a passive internet consumer, a "prosumer" at best, which should be taken in consideration in web corpus construction and computer-mediated communication studies. If blogs still bridge a technological gap between HTML-enhanced CMC and CMC-enhanced Web pages (Herring et al., 2004), a typological gap exists between original and current studies as well as between users of a platform and users of a content management system.

7 Conclusion

The trade-off to gain metadata using focused downloads following strict rules seems to get enough traction to build larger web corpora, since a total of 550 Gb of actually downloaded material allows after processing for the construction of a corpus of about 2.7 billion tokens with rich metadata. This comparatively high yield is a step towards more efficiency with respect to machine power and "Hi-Fi" web corpora, which could help promoting the cause of web sources and modernization of research methodology.

The resulting corpus complies with formal requirements on metadata-enhanced corpora and on weblogs considered as a series of dated entries. The interlinking of blogs and their rising popularity certainly don't stay in the way. However, addressing the tricky question of web genres seems inevitable in order to be able to properly qualify my findings and subsequent linguistic inquiries. More than ever, blogs are a hybrid genre, and their

ecology tends to mimic existing text types, audiences, and motivations, with a focus on information (general, specialized, or community-based) as well as on promotional goals.

References

Adrien Barbaresi and Kay-Michael Würzner. 2014. For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In *KONVENS 2014, NLP4CMC workshop proceedings*, pages 2–10. Hildesheim University Press.

Adrien Barbaresi. 2013. 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. 2014. Finding Viable Seed URLs for Web Corpora: A Scouting Approach and Comparative Study of Available Sources. In *Proceedings of the 9th Web as Corpus Workshop*, pages 1–8.

Adrien Barbaresi. 2015. *Ad hoc and general-purpose corpus construction from web sources*. Ph.D. thesis, École Normale Supérieure de Lyon.

Marco Baroni and Silvia Bernardini. 2004. BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of LREC*.

Rebecca Blood. 2002. *The Weblog Handbook: Practical Advice on Creating and Maintaining your Blog*. Basic Books.

Valter Crescenzi, Giansalvatore Mecca, and Paolo Merialdo. 2001. Roadrunner: Towards Automatic Data Extraction From Large Web Sites. In *Proceedings of the 27th VLDB Conference*, pages 109–118.

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.

Natalie Glance, Matthew Hurst, and Takashi Tomokiyo. 2004. Blogpulse: Automated Trend Discovery for Weblogs. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*.

Daniel Gruhl, Ramanathan Guha, David Liben-Nowell, and Andrew Tomkins. 2004. Information Diffusion through Blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501. ACM.

Yan Guo, Huifeng Tang, Linhai Song, Yu Wang, and Guodong Ding. 2010. ECON: an Approach to Extract Content from Web News Page. In *Proceedings of 12th International Asia-Pacific Web Conference (APWEB)*, pages 314–320. IEEE.

Ivan Habernal, Omnia Zayed, and Iryna Gurevych. 2016. C4corpus: Multilingual web-size corpus with free license. In *Proceedings of LREC*, pages 914–922.

Susan C Herring, Lois Ann Scheidt, Sabrina Bonus, and Elijah Wright. 2004. Bridging the Gap: A Genre Analysis of Weblogs. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, pages 11–21. IEEE.

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

Andrew Kehoe and Matt Gee. 2012. Reader comments as an aboutness indicator in online texts: introducing the Birmingham Blog Corpus. *Studies in Variation, Contacts and Change in English*, 12.

Sandeep Krishnamurthy. 2002. The Multidimensionality of Blog Conversations: The Virtual Enactment of September 11. *Internet Research*, 3.

Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. 2003. On the Bursty Evolution of Blogspace. In *Proceedings of WWW 2003*, pages 568–576.

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.

Craig Macdonald and Iadh Ounis. 2006. The TREC Blogs06 Collection: Creating and Analysing a Blog Test Collection. Technical report, Department of Computer Science, University of Glasgow.

Gilad Mishne and Natalie Glance. 2006. Leave a Reply: An Analysis of Weblog Comments. In *Third Annual Workshop on the Weblogging Ecosystem, WWW 2006*.

Tomoyuki Nanno, Toshiaki Fujiki, Yasuhiro Suzuki, and Manabu Okumura. 2004. Automatically Collecting, Monitoring, and Mining Japanese Weblogs. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 320–321. ACM.

Christopher Olston and Marc Najork. 2010. Web Crawling. *Foundations and Trends in Information Retrieval*, 4(3):175–246.

Michal Ptaszynski, Pawel Dybala, Rafal Rzepka, Kenji Araki, and Yoshio Momouchi. 2012. YACIS: A five-billion-word corpus of Japanese blogs fully annotated with syntactic and affective information. In *Proceedings of The AISB/IACAP World Congress*, pages 40–49.

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 and Felix Bildhauer. 2013. *Web Corpus Construction*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

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

Roland Schäfer, Adrien Barbaresi, and Felix Bildhauer. 2014. Focused Web Corpus Crawling. In *Proceedings of the 9th Web as Corpus workshop (WAC-9)*, pages 9–15.

Roland Schäfer. 2016. CommonCOW: massively huge web corpora from CommonCrawl data and a method to distribute them freely under restrictive EU copyright laws. In *Proceedings of LREC*, pages 4500–4504.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of Age and Gender on Blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205.

Vít Suchomel and Jan Pomikálek. 2012. Efficient Webcrawling for large text corpora. In *Proceedings of the 7th Web as Corpus Workshop*, pages 40–44.

Yannick Versley and Yana Panchenko. 2012. Not just bigger: Towards better-quality Web corpora. In *Proceedings of the 7th Web as Corpus Workshop*, pages 44–52.

Junfeng Wang, Xiaofei He, Can Wang, Jian Pei, Jiajun Bu, Chun Chen, Ziyu Guan, and Gang Lu. 2009. News Article Extraction with Template-Independent Wrapper. In *Proceedings of the WWW 2009*, pages 1085–1086. ACM.

Cai-Nicolas Ziegler and Michal Skubacz. 2007. Content Extraction from News Pages using Particle Swarm Optimization on Linguistic and Structural Features. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 242–249. IEEE.