The Vast and the Focused: On the need for thematic web and blog corpora

Adrien Barbaresi
Berlin-Brandenburg Academy of Sciences
Jägerstraße 22/23
D-10117 Berlin
barbaresi@bbaw.de

Abstract

As the Web ought to be considered as a series of sources rather than as a source in itself, a problem facing corpus construction resides in meta-information and categorization. In addition, we need focused data to shed light on particular subfields of the digital public sphere. Blogs are relevant to that end, especially if the resulting web texts can be extracted along with metadata and made available in coherent and clearly describable collections.

1 Problem description

The Web brings an unparalleled and rapidly evolving diversity in terms of speakers and settings. As such it should be considered as a series of sources rather than as a source in itself. Science needs an agreed scheme for identifying and registering research data (Sampson, 2000), in that sense schemes and methods are needed to live up to the potential of these potential sources for corpus construction. "Offline corpora" accessible within or throughout institutions are now standard among the research community. The process notably involves "crawling, downloading, 'cleaning' and deduplicating the data, then linguistically annotating it and loading it into a corpus query tool." (Kilgarriff, 2007) It relies on the assumption that "the Web is a space in which resources are identified by Uniform Resource Identifiers (URIs)." (Berners-Lee et al., 2006) The Web is however changing faster than the researchers' ability to observe it (Hendler et al., 2008), and a constant problem faced by web resources resides in meta-information and categorization. Due to the "heteregeneous and somewhat intractable character of the Web" (Bergh and Zanchetta, 2008), the actual contents of a web corpus can only be listed with certainty once the corpus is complete. In addition, web corpora exemplify "problems of large corpora built in short time and with little resources." (Baroni and Ueyama, 2006)

In fact, corresponding to the potential lack of information concerning the metadata of the texts is a lack of information regarding the content, whose adequacy, focus and quality has to be assessed in a post hoc evaluation (Baroni et al., 2009). The ability to describe a corpus accurately significantly increases its interest for researchers in the humanities and beyond. This is neither a trivial task nor a secondary one, as some assume that "text category is the most important organizing principle of most modern corpora" (O'Keeffe and McCarthy, 2010). Renouf (2007) also claims that lack of metadata makes an exhaustive study impossible or at least undermines it. Categories such as audience, authorship and artifact (Warschauer and Grimes, 2007), or authorship, mode, audience, aim, domain, and the annotation of textual dimensions (Sharoff, 2018) target this issue in particular.

Besides, a major fault line exists for the linguistic community between general and specific corpora (Gries, 2009). Since web corpora mostly follow from the existing linguistic tradition, their purpose and their methodology can also be divided into two main categories (Barbaresi, 2015). On the one hand there are all-purpose, "one size fits all" corpora, often designed to be large and diverse. On the other, there are specific corpora with controlled text inclusions and possibly rich metadata, built with particular research goals in mind, such as online news corpora or variation-aware approaches which take production conditions into account. This distinction also overlaps with diverging uses for corpora, for example corpus-based studies observing already known phenomena, and more opportunistically-minded research settings where size and content diversity allow for better coverage and use of statistical indicators. The contrast between general-purpose and specific corpora is not clear-cut as these categories are not impermeable: it is possible to find corpora that are in-between, or transferred from one to another due to later developments in corpus design.

2 From the vast to the focused

Seen from a practical perspective, the purpose of focused web corpora is to complement existing collections, as they allow for better coverage of specific written text types and genres, especially the language evolution seen through the lens of user-generated content, which gives access to a number of variants, socio- and idiolects. Methods consisting of "manually selecting, crawling and cleaning particular web sites with large and good-enough-quality textual content" (Spoustová and Spousta, 2012) are part of focused corpora, while focused crawling does not necessarily involve scrupulous work *a priori* but in any case the prioritization "towards documents which, according to some metric, have a high relevance" (Biemann et al., 2013). Even for comparatively large corpora, focused web corpus construction using pre-selected sources can lead to a higher yield and save time and resources while increasing the text quality of the resulting corpus (Schäfer et al., 2014).

The present use case concerns German, for which historical and contemporary corpora have been built as part of an aggregated lexical information platform (Geyken et al., 2017), the Digital Dictionary of the German Language (DWDS).[1] Specialized web corpora are built (Barbaresi, 2016) which can then be compared to existing resources such as newspaper and general-purpose corpora. Among other things, such corpora can be used to search for definitory elements related to newly created words or word senses (Barbaresi et al., 2018), for example by means of an automated content extraction and manual screening of pre-selected results.

A fundamental argument in favor of such corpora is related to the principles of the "Net economy" with the re-composition of the media landscape it fosters. It has seen the raise of "immaterial labor", "a social power that is independent and able to organize both its own work and its relations with business entities", where notions of "leisure time" and "working time" are fused and where the "split between author and audience" is transcended (Lazzarato, 1996). In some contexts the notion of "free labor" is also relevant to describe "the moment where [the] knowledgeable consumption of culture is translated into productive activities." (Terranova, 2000) These conditions of text production have to be accounted for, notably because they help creating a "long tail of bloggers who get little or no remuneration" (Rocamora, 2018) Community-building and content publishing among producers-consumers result in a major increase of text production which leads to more efficient corpus construction and potentially to a text collection that is easier to categorize.

Blogs seem to be particularly adequate as "the practice of blogging involves producing digital content with the intention of sharing it asynchronously with a conceptualized audience." (boyd, 2006) From the beginning of research on blogs/weblogs, the main definitory criterion has been their form, a reverse chronological sequences of dated entries and/or the use of dedicated software to articulate and publish the entries, a "weblog publishing software tool" (Glance et al., 2004) or content management system. Blogs are dynamic in nature, in consequence they "differ from static webpages because they capture ongoing expressions, not the edits of a static creation." (boyd, 2006) Another potential advantage in the case of focused crawls consists of the community-building aspects, as blogs are intricately intertwined in what has been called the blogosphere, as the active cross-linking helps to "create a strong sense of community" (Glance et al., 2004), which could help to find series of texts on a given topic by following links, that is by way of web crawling (Olston and Najork, 2010).

Difficulties raised by blogs as research objects are of conceptual and practical nature. First, the definition of what belongs to the genre and its use as a single category is controversial (Garden, 2012). This typology has notably been criticized for not being specific enough, especially concerning the sociolinguistic setting (Lomborg, 2009). A further demarcation can be made between blogs and social networks restricted to a single platform: "They differ from community tools because the expressions are captured locally, not in a shared common space." (boyd, 2006) These local spaces feature much less restrictions for machine-based access but also feature less directly exploitable

[1] https://www.dwds.de

metadata, although the profusion of user data on social media platforms can be of great value, for example to study linguistic variation (Barbaresi and Ruiz Tinoco, 2018). Consequently, the extraction of relevant content and metadata is highly relevant in order to make such web corpora exploitable. Finally, the commonly found term of blogosphere suggests a connection that does not necessarily exist, in opposition to the concept of "blogipelago", which "reminds us of separateness, disconnection, and the immense effort it can take to move from one island or network to another" (Dean, 2010). This effort clearly impacts corpus construction by requiring more screening as well as significant "island hopping". This is for example the case in communities which are fairly small and disconnected from other websites on the topic, e.g. Austrian fashion blogs which appear to refer to each other but do not often include links to other similar communities or topics. In the end, it is quite rare to find ready-made resources, especially for a topically focused approach, so that gathering methods and criteria ought to be discussed. As in genre-based studies, manual annotation – for example through crowdsourcing – can be an option for assessing the content of web texts and pave the way for classification tasks, but the lack of pre-existing data makes a pioneering work necessary (Asheghi et al., 2014). Provided this assumption is correct, collecting restricted portions of the Web for linguistic research remains nevertheless possible with sufficient screening.

3 Preliminary conclusions

Following the research on blogs/weblogs, we define blogs according to their form, consisting of dated entries available online and often managed by a broadly available publishing tool or web space. The discovery of relevant portions of the web is performed semi-automatically by pre-selecting hundreds of sources. Second, important metadata such as the publication date and main text content are extracted automatically based on structural patterns as well as heuristic criteria on text and markup. The resulting text base resides in a subset of web pages which have been found, downloaded and processed; documents with non-existent or missing date or entry content are discarded during processing and are not part of the corpus. By checking the seen web pages as to their relevance, it becomes possible to benefit from the insertion into a "web territory" (Cardon et al., 2011) that implies virtual communities as well as a complex adaptation process, which is also relevant from a linguistic standpoint. Surveys of particular portions of the web can also feature additional criteria such as content licensing, as some public licenses could help contributing back the corpus construction work to the research community.

We need both data and scientific instruments to shed light on subfields of the digital public sphere such as websites devoted to information technology (Pohlmann and Barbaresi, 2019), fashion & beauty, or literature. These topics in particular have the advantage of being among the most present online while mostly addressing complementary "prosumer" communities, even if studies relying on website publishing and blogging activities face a long tail with respect to impact and readership as well as concerning the move towards other publishing platforms and other content types. Nonetheless, some interlinking exists, webpages and especially blogs are still alive and relevant to gather corpus evidence. In the end, compared to "pre-web" and general-purpose corpora, challenges reside (1) in the necessity to consider texts types and topics beyond the previous extension of these notions and beyond known categories, (2) in a corresponding mapping of relevant portions of the web, and (3) in the ability to extract and pre-process resulting web texts and ultimately to make them available in clearly describable and coherent collections.

References

Noushin Rezapour Asheghi, Serge Sharoff, and Katja Markert. 2014. Designing and Evaluating a Reliable Corpus of Web Genres via Crowd-Sourcing. In *9th Language Resources and Evaluation Conference (LREC 2014)*, pages 1339–1346.

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

Adrien Barbaresi. 2016. Efficient construction of metadata-enhanced web corpora. In *Proceedings of the 10th Web as Corpus Workshop*, pages 7–16. Association for Computational Linguistics.

Adrien Barbaresi, Lothar Lemnitzer, and Alexander Geyken. 2018. A database of German definitory contexts from selected web sources. In *11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3068–3073. European Language Resources Association (ELRA).

Adrien Barbaresi and Antonio Ruiz Tinoco. 2018. Using Elasticsearch for Linguistic Analysis of Tweets in Time and Space. In *Proceedings of the LREC 2018 Workshop CMLC-6*, pages 14–19. ELRA.

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 and Motoko Ueyama. 2006. Building general- and special-purpose corpora by Web crawling. In *Proceedings of the 13th NIJL International Symposium, Language corpora: Their compilation and application*, pages 31–40.

Gunnar Bergh and Eros Zanchetta. 2008. Web linguistics. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics, An International Handbook*, pages 309–327. Mouton de Gruyter, Berlin.

Tim Berners-Lee, Wendy Hall, James A. Hendler, Kieron O'Hara, Nigel Shadbolt, and Daniel J. Weitzner. 2006. A Framework for Web Science. *Foundations and Trends in Web Science*, 1(1):1–130.

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.

danah boyd. 2006. A Blogger's Blog: Exploring the Definition of a Medium. *Reconstruction*, 6(4):1–21.

Dominique Cardon, Guilhem Fouetillou, and Camille Roth. 2011. Two Paths of Glory – Structural Positions and Trajectories of Websites within Their Topical Territory. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Jodi Dean. 2010. *Blog theory: Feedback and capture in the circuits of drive*. Polity, Cambridge.

Mary Garden. 2012. Defining blog: A fool's errand or a necessary undertaking. *Journalism*, 13(4):483–499.

Alexander Geyken, Adrien Barbaresi, Jörg Didakowski, Bryan Jurish, Frank Wiegand, and Lothar Lemnitzer. 2017. Die Korpusplattform des "Digitalen Wörterbuchs der deutschen Sprache" (DWDS). *Zeitschrift für germanistische Linguistik*, 45(2):327–344.

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*, volume 2004, New York.

Stefan T. Gries. 2009. What is Corpus Linguistics? *Language and Linguistics Compass*, 3(5):1225–1241.

James Hendler, Nigel Shadbolt, Wendy Hall, Tim Berners-Lee, and Daniel Weitzner. 2008. Web Science: An Interdisciplinary Approach to Understanding the Web. *Communications of the ACM*, 51(7):60–69.

Adam Kilgarriff. 2007. Googleology is bad science. *Computational Linguistics*, 33(1):147–151.

Maurizio Lazzarato. 1996. Immaterial Labor. In P. Virno and M. Hardy, editors, *Radical Thought in Italy*, pages 132–146. University of Minnesota Press.

Stine Lomborg. 2009. Navigating the blogosphere: Towards a genre-based typology of weblogs. *First Monday*, 14(5).

Anne O'Keeffe and Michael McCarthy. 2010. *The Routledge handbook of corpus linguistics*. Routledge.

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

Jens Pohlmann and Adrien Barbaresi. 2019. Diving into the Conplexities of the Tech Blog Sphere. In *Digital Humanities 2019 Book of Abstracts*. ADHO.

Antoinette Renouf. 2007. Corpus development 25 years on: from super-corpus to cyber-corpus. In *Corpus Linguistics 25 years on*, pages 27–49. Brill Rodopi.

Agnès Rocamora. 2018. The labour of fashion blogging. In Leah Armstrong and Felice McDowell, editors, *Fashioning Professionals*. Bloomsbury.

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. 2014. Focused Web Corpus Crawling. In *Proceedings of the 9th Web as Corpus workshop (WAC-9) @ EACL 2014*, pages 9–15. Association for Computational Linguistics.

Serge Sharoff. 2018. Functional text dimensions for the annotation of web corpora. *Corpora*, 13(1):65–95.

Johanka Spoustová and Miroslav Spousta. 2012. A High-Quality Web Corpus of Czech. In *Proceedings of LREC*, pages 311–315.

Tiziana Terranova. 2000. Free labor: Producing culture for the digital economy. *Social text*, 18(2):33–58.

Mark Warschauer and Douglas Grimes. 2007. Audience, authorship, and artifact: The emergent semiotics of web 2.0. *Annual Review of Applied Linguistics*, 27:1–23.