Diving Into The Complexities Of The Tech Blog Sphere

Jens Pohlmann, Adrien Barbaresi

▶ **To cite this version:**

Jens Pohlmann, Adrien Barbaresi. Diving Into The Complexities Of The Tech Blog Sphere. Digital Humanities 2019, ADHO, Jul 2019, Utrecht, Netherlands. hal-02201532

HAL Id: hal-02201532
https://hal.science/hal-02201532

Submitted on 31 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

Diving Into The Complexities Of The Tech Blog Sphere

XML

Following the assumption that the tech blog sphere represents an avant-garde of technologically and socially interested experts, we describe an experimental setting to observe its input on the public discussion of matters situated at the intersection of technology and society. Our interdisciplinary approach consists in joining forces on a common base of texts and tools. This cooperation stems from work on the impact of digital media on democratic processes and institutions (GHI/RRCHNM) and corpus and computational linguistics (BBAW). The major aims of the effort described here are twofold: (1) compiling a text base (for German and English) from a curated list of blogs dedicated to technological topics for lexicographical and linguistic research, as well as (2) conducting exemplary studies using the compiled corpus, focusing on specific research questions regarding public discourse in Germany and the United States on questions of internet policy.

1. A Tech Blog Corpus for Linguistic Research and Discourse Analysis

The purpose of focused web corpora in linguistic research 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, especially for a topically centered approach. Blogs are of particular interest for our research, since they are intricately intertwined in what has been called the blogosphere, as the active cross-linking helps “create a strong sense of community” (Glance 2004). Specific tech blogs first evolved aside from and in opposition to traditional mass media settings and amateur blogs have been shown to have the capacity to open up public space for the debate of socially relevant issues. Technological questions are indeed not restricted to the world of specialists, precisely since their implications often turn into political and ethical realities that affect society as a whole. However, the small, local communities of the beginnings have mostly been relegated by commercially driven websites targeting passive readers, which certainly has an impact on the content of discussions. These tech blog outlets may thereby have taken on the status of influential information providers, agenda setters, and gatekeepers.

We need both data and scientific instruments to shed light on this subfield of the digital public sphere. Our first use case focuses on 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). For the project presented here, we construct a specialized web corpus (Barbaresi 2016) which will then be compared to existing generalist resources such as newspapers and websites. Following the research on blogs/weblogs, we define blogs according to their form, consisting of dated entries 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 a series of heuristics. The resulting text base resides in a subset of web pages which could be found, downloaded and processed; documents with non-existent or missed date or entry content are discarded during processing and are not part of the corpus.

2. Application and Case Studies

The corpus is used (1) to search for definitory elements related to newly created words or word senses (Barbaresi et al. 2018), which involves an automated extraction of content and manual screening, and (2) to study discussions on lawmaking, which involves finding one’s way through convoluted and heterogenous documents, a task for which philologists can be assisted by large specialized text corpora and databases as well as distant reading processes such as topic modeling and outlier detection. One exemplary topic in the examination of this corpus focuses on the public discussion of the German Network Enforcement Act or “NetzDG”. This controversial

anti-hate speech law, which forces social media platforms to take down flagged content that is “manifestly unlawful” within 24 hours of receiving the complaint, has been discussed and criticized in Germany, but has been very much condemned in the United States. The criticism put forward focuses on the abetting of overblocking that may lead to forms of censorship, on the outsourcing of juridical decisions to private companies, and on setting examples for authoritarian regimes’ copycat laws. The discourse about “NetzDG” is an extremely relevant case study for the analysis of the ways in which the societal implications of technology are currently discussed and negotiated in the public sphere, especially with regards to the threats that the misuse of social media platforms poses to political decision-making processes in Western democracies. The discourse about “NetzDG” particularly points to the diverging cultures regarding freedom of expression in Germany and the United States and it illustrates the extent to which the historical roots of these differences inform the current transatlantic debate about the restriction of content online and the regulation of social media platforms (Nieuwenhuis 2000, Schulz, 2018). The debate in itself includes a highly technical vocabulary and the need to transfer knowledge from a small community of experts to the general public. Finally, we also tackle questions related to the changing nature of the Web, as a web corpus almost instantly turns into a web archive, which calls for the long-term examination of its content through philological concepts and technical tools which are crafted expressly.

3. Further Developments

This specially compiled blog corpus will shed light on processes of societal negotiation located at the crossroads of technology, public policy, society and the Internet, most notably by providing access to discourses in the tech blog sphere and by allowing for comparisons with other text types and sources. Aside from investigations regarding free speech issues and the “NetzDG,” potential inquiries include topics such as privacy laws, upload filters, AI, or copyright legislation in the digital age, as well as the analysis of the communication strategies employed by the respective stakeholders and the linguistically distinct characteristics observed. We plan to make a series of resources available in order to support the cause of web sources and the modernization of research methodology. The tools we work with or develop are released under open-source licenses and our sources will be published as curated lists of websites. In addition, the texts will be included into the DWDS web platform so that they can be queried by the wider public. If and to the extent applicable, we will release corpus data to apply and test further methods of automated text analysis.

Appendix A

Bibliography

1. **Barbaresi, A.** (2016). Efficient construction of metadata-enhanced web corpora. In Proceedings of the 10th Web as Corpus Workshop, Association for Computational Linguistics (ACL SIGWAC), 7-16.
2. **Barbaresi, A.**, **Lemnitzer, L. & Geyken, A.** (2018). A database of German definitory contexts from selected web sources. Proceedings of LREC 2018, ELRA, 3068-3073.
3. **Geyken, A., Barbaresi, A., Didakowski, J., Jurish, B., Wiegand, F., & Lemnitzer, L.** (2017). Die Korpusplattform des „Digitalen Wörterbuchs der deutschen Sprache “(DWDS). Zeitschrift für germanistische Linguistik, 45(2), 327-344.
4. **Glance, N., Hurst, M., & Tomokiyo, T.** (2004). Blogpulse: Automated trend discovery for weblogs. In : *WWW 2004 workshop on the weblogging ecosystem: Aggregation, analysis and dynamics*.
5. **Nieuwenhuis, A.** (2000). Freedom of Speech: USA vs. Germany and Europe. Netherlands Quarterly of Human Rights, 18(2), 195–217.
6. **Schulz, W.** (2018). Regulating Intermediaries to Protect Privacy Online – The Case of the German NetzDG (SSRN Scholarly Paper No. ID 3216572). Rochester, NY: Social Science Research Network.

Jens Pohlmann (jop@stanford.edu), German Historical Institute (GHI), Washington DC, United States of America; Roy Rosenzweig Center for History and New Media, George Mason University, USA und Adrien Barbaresi (barbaresi@bbaw.de), Berlin-Brandenburg Academy of Sciences (BBAW), Germany