Collection, Description, and Visualization of the German Reddit Corpus

Adrien Barbaresi
Austrian Academy of Sciences (ÖAW)
Berlin-Brandenburg Academy of Sciences (BBAW)
adrien.barbaresi@oeaw.ac.at

Abstract

Reddit is a major social bookmarking and microblogging platform. An extensive dataset of Reddit comments has recently been made publicly available. I use a two-tiered filter to single out comments in German in order to build a linguistic corpus which is then tokenized and annotated. This article offers first insights of both nature and quality of data at the lexical level. Additionally, a visualization makes it possible to grasp the possible geographical distribution of German users of the platform.

1 Introduction

One of the main issues when dealing with web corpora, be it general-purpose corpora or specific ones, consists in the discovery of relevant web documents for linguistic studies. There are for example few projects dealing with computer-mediated communication in German, and it is quite rare to find ready-made resources. The DeRiK project for instance features ongoing work with the purpose to build a reference corpus dedicated to computer-mediated communication (Beißwenger et al., 2013). Previous work towards the constitution of a German blog corpus under CC license implied a significant effort (Barbaresi and Würzner, 2014).

In this respect, it has been particularly surprising to hear from the release of a complete dataset of comments published on Reddit, a major social network. This article describes the steps taken in order to get a first glimpse of German data in the corpus as well as to describe what makes CMC-data in general and Reddit data in particular so different.

One hope is that the Reddit corpus can be used to find relevant examples of previously undocumented language uses for lexicography and dictionary building projects, e.g. the DWDS lexicography project (Geyken, 2007), and/or to test linguistic annotation chains for robustness.

2 Description of the dataset

2.1 Reddit

Reddit is a social bookmarking and microblogging platform owned by the American mass media company Condé Nast. It ranks at first place worldwide in the news category according to the site metrics aggregator Alexa[1], which makes it a typical Internet phenomenon. The short description of the website according to Alexa is as follows: "User-generated news links. Votes promote stories to the front page." Indeed, the entries are organized into areas of interest called "reddits" or "subreddits", which are curated by the users ("redditors") themselves. Since the moderation processes are mature, and since the channels (or subreddits) have to be hand-picked, they ensure a certain stability. From a linguistic point of view, one may say that users account for the linguistic homogeneity if not relevance of their channel.

There is an API for Reddit, allowing automated retrieval of comments. However, search depth is limited: it is often not possible to go back in time further than the 500th oldest post, which severely restricts the number of links one may crawl (Barbaresi, 2013). Continuous crawling is then necessary in order to gather all the possible comments on all the subreddits.

2.2 Original release

The work described in the article directly follows from the recent release of the "Reddit comment corpus": Reddit user *Stuck_In_The_Matrix* (Jason Baumgartner) made the dataset publicly available on the platform archive.org[2] at the beginning of July 2015. In its original release statement on Reddit, Baumgartner claims to have gathered every publicly available Reddit comment, which amounts

[1] http://www.alexa.com/topsites/category/Top/News
[2] https://archive.org/details/2015_reddit_comments_corpus

to 1.65 billion JSON objects.[3] 350,000 comments out of 1.65 billion were unavailable due to Reddit API issues.

While its compiler chose to name it a "corpus", the whole could rather be called a dataset. In fact, apart from ensuring the most complete collection process possible, no specific steps were taken to allow for a control of the contents in the sense of the linguistic tradition (Barbaresi, 2015).

2.3 Filtering steps

I use a two-tiered filter in order to deliver a hopefully well-balanced performance between speed and accuracy. The combined strategy proved efficient in preliminary tests as well as in previous studies (Lui and Baldwin, 2014). The first filter consists in a dictionary-based approach taking benefit from spell-checking algorithms. It discriminates between comments using thresholds expressed as a percentage of tokens which do not pass the spell check. The second filter is a full-fledged language detection software, which outputs the most probable language according to its model.

First, spell checking algorithms benefit from years of optimization concerning both speed and accuracy. The library used, *enchant*, allows the use of a variety of spell-checking backends, like aspell, hunspell or ispell, with one or several locales.[4] English being the most prominent language on Reddit, each token is tested for errors in both English and German. A comment which induces a relatively high amount of errors for English (more than 30%) but a relatively low one for German (less than 70%) is considered to be interesting enough to proceed to the second step. In other studies (Barbaresi, 2013), I have used a threshold of 0.5; while I did not witness significant changes on Reddit data, I still chose a more defensive setting in order to ensure corpus relevance.

Second, a language identification tool is used to maximize the precision of the language recognition. `langid.py` (Lui and Baldwin, 2012) is open-source[5], it incorporates a model which has been pre-trained on a variety of web documents (Clue Web and Wikipedia inter alia). It has already been used to classify social media text on a large scale (Baldwin et al., 2013) and it is fast enough to be able to classify data from the order of magnitude of the Reddit comments.

The filtering described here is reproducible and can be attempted using other parameters, instructions and code to do so are available.[6]

3 Analysis of the corpus

3.1 Linguistic features

The corpus has been tokenized by WASTE (Jurish and Würzner, 2013) and lemmatized by MOOT (Jurish, 2003). It contains a total of 97,505 comments, 89,681 sentences, 566,362 tokens, and 3,352,472 characters. It is clear that Reddit is almost exclusively an English-speaking platform, however there are eminent German channels and due to the sheer size of the original dataset one could have expected a larger corpus. Maybe the precision of the filters could be lowered in favor of a better recall.

The mean token and sentence length (respectively 5.92 characters and 6.32 tokens) are in line with the expectations concerning computer-mediated communication, and it clearly anchors the corpus on this side of the spectrum. The relatively large vocabulary size in terms of types with 64,314 different forms (27% of which are hapax legomena) calls for further analyses. Qualitatively speaking, the ironic tone Reddit is known for could also prove to be interesting.

POS-tag	Frequency
NN	17.6%
NE	14.4%
ADV	8.6%
VVFIN	6.8%
PPER	5.9%
ADJD	4.7%
ART	4.3%
VAFIN	4.1%
XY	3.8%
ADJA	3.5%

Table 1: Most frequent POS-tags and relative frequency on token level, without spaces and punctuation

The breakdown into different part-of-speech tags shown in table 1 gives insights on the actual contents. The proportion of a number of tags from the STTS tagset is in line with other general or CMC-corpora (Barbaresi, 2014), however the number of tokens tagged as proper nouns (NE) is particularly

[3] https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/

[4] http://www.abisource.com/projects/enchant/

[5] https://github.com/saffsd/langid.py

[6] https://github.com/adbar/german-reddit

high (14.4%), which exemplifies the perplexity of the tool itself, for example because the redditors refer to trending and possibly short-lived notions and celebrities, or because of a high proportion of short, elliptic comments which fail to provide enough morpho-syntactic context. The relatively high but acceptable proportion of foreign words on token level (4%) both confirms this hypothesis and validates the language classification performed during corpus building.

Smiley	Frequency	Smiley	Frequency
:)	3207	:-(	39
;)	1667	-.-	33
:(	590	:'(	31
:-)	299	:))	29
^^	242	:]	25
;-)	238	=(	19
:/	179	:|	18
=)	96		

Table 2: Most frequent smileys and their frequency

Thanks to the special training of the tokenizer on CMC-data, the smileys can be expected to be treated as whole tokens, which makes a focused analysis possible. The major part of frequent smileys listed in table 2 is commonly used, although there are idioms such as "=)" which may be more frequent on this platform. Emojis do not seem to be frequently used in German comments.

3.2 Sociolinguistic factors

Information about the subreddit of each comment is part of the JSON metadata, which makes the extraction of subreddits straightforward. As can be seen in table 3, the most frequent ones include channels where expression in German is expected (*de*, *rocketbeans*, *kreiswichs*) and other where German is not necessarily spoken but could be appropriate (*germany*, *Austria*). Other channels, which are known to be among the most popular ones but whose link to German is not clear, may include enough quotes or occasional discussions (e.g. *AskReddit*) to explain their presence among the most frequent ones.

The nicknames are also part of the metadata returned by the API and as such they can be considered to be reliable information. A total of 51,155 different nicknames can be found throughout the German subset. 5,343 are marked as deleted, i.e. not active at the time of download.

The most frequent author names in table 4 show

Channel	Frequency
de	14018
AskReddit	8163
rocketbeans	4899
funny	3272
kreiswichs	2848
pics	2813
soccer	2571
Austria	1684
WTF	1592
leagueoflegends	1569
reddit.com	1379
todayilearned	1224
germany	1137
gaming	1133
videos	1124

Table 3: Most frequent channels (subreddits) in the corpus

that although there is a slight trend towards typical German nouns or syllables, they are not the majority. The crowd seems to be relatively evenly distributed, there is no single nickname outweighing all the others. That said, it can be common practice to change nicknames regularly, which also accounts for the relatively high number of deleted accounts found.

Nickname	Frequency
Wumselito	262
Aschebescher	238
Clit_Commander	221
oldandgreat	210
GuantanaMo	200
fLekkZ	187
Obraka	180
tin_dog	155
4-jan	151
Omnilatent	141

Table 4: Most frequent nicknames in the corpus and their frequency

4 Visualization of extracted place names

4.1 Method

The Reddit comments are not geotagged. Thus, a proxy has to be found in order to get a glance at their socio-geographical distribution. To do so, place names are extracted and projected on a map,

which allows for a better description of the collected data.

First, the German version of the Wiktionary, a user-curated dictionary launched by the Wikimedia foundation, is used in order to get lexical information about common nouns, which allows for a fine-grained discriminating analysis.

Second, geographical information about the places names has been compiled from the Geonames database[7], which is e.g. used by the Openstreetmap project[8], and whose Creative Commons Attribution license will allow for a release of research data in the near future. All databases for current European countries have been retrieved and preprocessed certain place types have been selected. In fact, toponym resolution often relies on named-entity recognition and artificial intelligence (Leidner and Lieberman, 2011), but knowledge-based methods using fine-grained data have already been used with encouraging results (Hu et al., 2014).

The tokenized corpus has been filtered as described above and matched with the database. This operation includes finite-state automatons at two distinct stages: first to discover potential multiword place names, and second to select the most probable coordinates in the case of homonyms, based on type, relative distance, and population.

4.2 Results

The maps in figure 1 have been generated by the design environment TileMill[9] and customized using the stylesheet language CartoCSS. Both maps were created using the same data, on the left the scale is smaller, while on the right place names for frequent entries have been added.

The linguistic corpora at the basis of the maps are a construct, and so are the maps themselves: although they seem immediately interpretable, the quality of data, the specialization of the processing tools, and quality assessment all have a major impact on the outcome.

The place names seem to be quite evenly distributed in the German language area relatively to the most populated cities and thus the expectations. There are a few interesting exceptions: Berlin is usually more precisely named (e.g. "Berlin-Kreuzberg" instead of just "Berlin"), which gives the capital the shape of a constellation on the map. A clustering phase could be necessary in order to be able to compare it to the other main cities.

All in all, cities the western part of Germany seems to be more frequently mentioned, particularly when they are home to a well-known soccer team (such as Mönchengladbach). In Austria, Vienna clearly outweighs the rest of the country, which could be explained by a higher international visibility as well as a higher density of early-adopters of Reddit.

5 Conclusion

In this article, I have shown how a corpus focusing on German can be built using the publicly available Reddit comment dataset. In order to get a first impression of the corpus, I collected quantitative information and offered a visualization of structured data, more precisely place names which have to be extracted from the comments since they are not geotagged.

The structural properties of the corpus are in line with the expectations concerning CMC (Barbaresi and Würzner, 2014): short sentences, a relatively high number of different lemmata, and a whole vocabulary of smileys. The different nicknames involved seem to be rather evenly distributed, so are the different place names mentioned in the comments, which is good news in terms of diversity.

Since the license restrictions concerning the dataset are unclear, the corpus is only available upon request. Nonetheless, the German subset can be reconstructed and updated from scratch using code released under open source license.[10]

References

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364.

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.

[7] http://www.geonames.org

[8] https://www.openstreetmap.org

[9] https://www.mapbox.com/tilemill/

[10] https://github.com/adbar/german-reddit

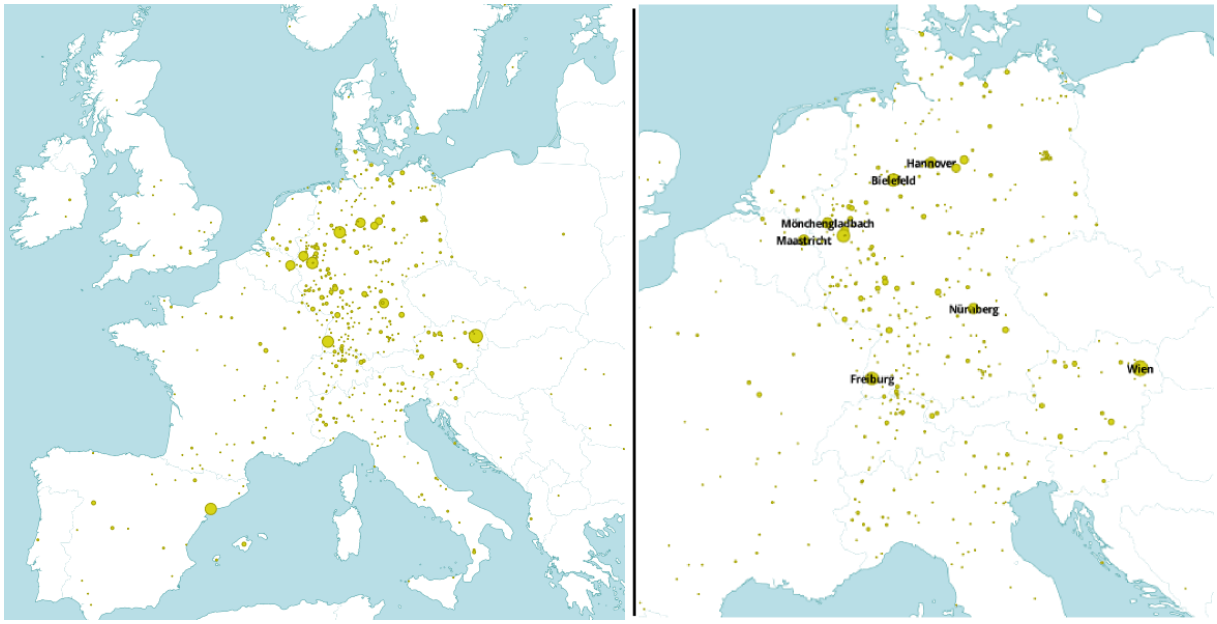

Figure 1: Projection of extracted place names on maps

Adrien Barbaresi. 2014. Finding Viable Seed URLs for Web Corpora: A Scouting Approach and Comparative Study of Available Sources. In Roland Schäfer and Felix Bildhauer, editors, *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.

Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Angelika Storrer. 2013. DeRiK: A German reference corpus of computer-mediated communication. *Literary and Linguistic Computing*, 28(4):531–537.

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.

Yingjie Hu, Krzysztof Janowicz, and Sathya Prasad. 2014. Improving Wikipedia-Based Place Name Disambiguation in Short Texts Using Structured Data from Dbpedia. In *Proceedings of the 8th Workshop on Geographic Information Retrieval*, pages 8–16. ACM.

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

Bryan Jurish. 2003. A Hybrid Approach to Part-of-Speech Tagging. Final report, Kollokationen im Wörterbuch, Berlin-Brandenburgische Akademie der Wissenschaften.

Jochen L Leidner and Michael D Lieberman. 2011. Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. *SIGSPATIAL Special*, 3(2):5–11.

Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju, Republic of Korea.

Marco Lui and Timothy Baldwin. 2014. Accurate Language Identification of Twitter Messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 17–25.