Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)

A New German Reddit Corpus

Andreas Blombach and **Natalie Dykes** and **Stefan Evert** and
Philipp Heinrich and **Besim Kabashi** and **Thomas Proisl**
Lehrstuhl für Korpus- und Computerlinguistik
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, Germany
{andreas.blombach,natalie.mary.dykes,stefan.evert}@fau.de
{philipp.heinrich,besim.kabashi,thomas.proisl}@fau.de

Abstract

We describe the creation of a German Reddit corpus, difficulties encountered along the way, and some of the data's linguistic peculiarities.

1 Data Gathering

What's Reddit? Reddit is a platform combining social news aggregation, discussion and microblogging. Since its founding in 2005, it has grown to be one of the most popular websites in the USA; in recent years, its popularity has also increased in Germany, as indicated by site rankings from Alexa and SimilarWeb.[1]

Users submit content (e.g. text, images or links) and can comment on submissions. Submissions and comments can be voted up or down, affecting the order in which they are displayed (submissions with the most upvotes make it to the front page).

Reddit is structured into so-called "subreddits" with their own community rules. Subreddits range from being rather open-topic (e.g. *r/de* – anything related to German) to extremely specific.

The German Reddit Sphere Just as subreddits' topics and contents vary widely, so do linguistic phenomena associated with particular subreddits. While some subreddits exhibit mostly standard language, others have rather unique memes and practices; making them difficult for outsiders to understand. In German subreddits, for instance, emoticons may be replaced by German *Umlaut* characters:

:) → Ü   :o → Ö   :< → Ä

On a lexical and phraseological level, typical expressions commonly associated with online communication as well as Reddit-specific ones are often translated word for word, leading to a humorous effect: *pfostieren* 'to post', *ausgelöst* 'triggered', *Unterlases* 'subreddit', *fixierte das für dich* 'fixed this for you'.

2 Corpus Creation

Building a Reddit Corpus In an ongoing effort, Jason Baumgartner collects every Reddit submission and comment, publicly accessible via `https://files.pushshift.io/reddit/`[2] (some caveats apply, see Gaffney and Matias (2018)). The first attempt at extracting German comments was made by Barbaresi (2015). At the time, Reddit was less widely used, especially by German-speaking users, and the resulting corpus was relatively small (97,505 comments, 566,362 tokens).

Detecting German Comments To detect German comments in the vast dataset, we adapt Barbaresi's approach of using a two-tiered filter relying on spell checking and language identification. After some sanitizing steps to ignore extremely short or deleted comments, every token is checked against a German and an English dictionary with a regular expression. A text is classified as potentially German if at least 70% of its tokens are found in the dictionary, and no more than 30% are present in the English dictionary. Next, `langid` (Lui and Baldwin, 2014) is run on these candidate comments to ultimately classify comments as German. This way, we identified more than 6,700,000 German comments between 2016 and 2018, amounting to roughly 230,000,000 tokens running text.

Evaluation of the Two-tiered Filter In a random sample of 1,618 comments, we manually identified 26% which are not actually in German. While longer comments are recognized rather reliably,

[1] Ranks as of October 9, 2019:
Alexa: 6 (US), 9 (DE); SimilarWeb: 10 (US), 31 (DE); see `https://www.alexa.com/topsites/countries` and `https://www.similarweb.com/top-websites`

[2] See also `https://www.reddit.com/r/pushshift/comments/bcxguf/new_to_pushshift_read_this_faq/`

Subreddit	“German”	Total	Fraction
de	3,644,481	4,364,440	0.835
Austria	389,041	483,125	0.805
rocketbeans	142,787	164,699	0.867
AskReddit	125,411	183,147,454	**0.000685**
edefreiheit	121,370	147,150	0.825

Table 1: Comment counts in the top 5 subreddits with “German” comments, 2016–2018

shorter comments are often misclassified – for instance, due to the presence of proper names.

Possible Corpus Cleanup Table 1 shows the top 5 subreddits containing comments identified as German. We interpret the low relative frequencies of German comments in some subreddits as an indication of false positives. Thus, filtering out subreddits with less than e. g. 10% of German comments seems like a plausible strategy. On the other hand, for many predominantly German subreddits, the two-tiered filter may have been too aggressive, and it might be sensible to retain all comments from these to keep conversations intact.[3] In any case, using the filter to identify predominantly German subreddits in the first place works as intended.

3 Annotation

Pre-processing and Tokenization Some of Reddit's peculiarities pose challenges to existing tools. Comments can include Markdown markup, making pre-processing necessary. Short form links to subreddits and user profiles follow the pattern *r/oldschoolcool* and *u/username*, which has to be accounted for by a tokenizer. Punctuation is often omitted and replaced by line breaks or emoticons: *So bin endlich zu was gekommen :D Habe [...]*

Tokenization and POS Tagging A random sample of comments was tokenized using SoMaJo (Proisl and Uhrig, 2016) and tagged with the STTS_IBK tagset (Beißwenger et al., 2015) using SoMeWeTa (Proisl, 2018). To evaluate performance, manual correction was performed on 1,186 tokens. The tagging accuracy was at 92.6%. POS errors seemed to be largely systematic, with the very fine-grained differentiation in particle types being hard to achieve due to sparseness in the training data, i.e. the EmpiriST corpus (Beißwenger et al., 2016).

Our evaluation leads us to the question whether a revised CMC tagset might be beneficial: While there are fine-grained categories for e.g. particle types, more obvious distinctions are not made (e.g. between definite and indefinite articles). Moreover, only certain contractions are assigned tags, and no differentiation is made for common acronyms (*scnr, imho*) and different types of punctuation (also affecting asterisks marking “action words” like **lol**).

4 Outlook

Due to their peculiarities, Reddit data are a promising source for further (socio-)linguistic research. Since the same features pose challenges to existing tools, more corpus cleaning will be necessary and rules for tokenization and tagging will need to be updated.

References

Adrien Barbaresi. 2015. Collection, description, and visualization of the German Reddit corpus. In *Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media*, pages 7–11, Essen.

Michael Beißwenger, Thomas Bartz, Angelika Storrer, and Swantje Westpfahl. 2015. Tagset und Richtlinie für das Part-of-Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation. Guideline document.

Michael Beißwenger, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. 2016. EmpiriST 2015: A shared task on the automatic linguistic annotation of computer-mediated communication and web corpora. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 44–56, Berlin. ACL.

Devin Gaffney and J. Nathan Matias. 2018. Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PLOS ONE*, 13(7):1–13.

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, Gothenburg, Sweden. ACL.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin. ACL.

Thomas Proisl. 2018. SoMeWeTa: A part-of-speech tagger for German social media and web texts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 665–670, Miyazaki. ELRA.

[3] Subreddits containing dialectal language use (such as *r/aeiou* or *r/BUENZLI*) are a special case.