A Corpus of German Reddit Exchanges (GeRedE)

Andreas Blombach, Natalie Dykes, Philipp Heinrich, Besim Kabashi, Thomas Proisl
Computational Corpus Linguistics Group
Friedrich-Alexander-Universität Erlangen-Nürnberg
Bismarckstr. 6, 91054 Erlangen, Germany
{andreas.blombach, natalie.mary.dykes, philipp.heinrich, besim.kabashi, thomas.proisl}@fau.de

Abstract

GeRedE is a 270 million token German CMC corpus containing approximately 380,000 submissions and 6,800,000 comments posted on Reddit between 2010 and 2018. Reddit is a popular online platform combining social news aggregation, discussion and micro-blogging. Starting from a large, freely available data set, the paper describes our approach to filter out German data and further pre-processing steps, as well as which metadata and annotation layers have been included so far. We explore the Reddit sphere, what makes the German data linguistically peculiar, and how some of the communities within Reddit differ from one another. The CWB-indexed version of our final corpus is available via CQPweb, and all our processing scripts as well as all manual annotation and automatic language classification can be downloaded from GitHub.

Keywords: Reddit, Corpus building, Corpus analysis, Social media, CMC, German

1. Introduction

Reddit, the self-proclaimed 'front page of the internet', combines social news aggregation, discussion and micro-blogging. 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]

Reddit is structured into so-called "subreddits" (categories or communities), whose moderators can define and enforce their own community rules (there is, of course, also a sitewide content policy). Subreddits range from being rather open-topic (e. g. *r/de* – anything related to German and Germany) to extremely specific (e. g. *r/wasletztepreis*, roughly translating to 'whatlastprice' – a sarcastic subreddit dedicated to negative interactions on the German version of Ebay Classified Ads).

Users can submit content (e. g. text, images or links) to a subreddit and comment on others' submissions and comments, resulting in nested conversation threads. Both submissions and comments can be voted up or down by the community. The items' voting score affects the default order in which content is displayed (submissions with the most upvotes are shown on the front page).

1.1. Userbase

Reddit's last official count of monthly active users from November, 2017 puts them at more than 330 million[2], a number comparable to that of Twitter. The vast majority of Reddit users (sometimes called "redditors") comes from English-speaking countries, especially the USA. With some caution, we can probably generalize some information we have on US users: while more women than men use social media in general[3], Reddit is much more popular among men than among women.[4] Users are also mostly young (with the highest share of adult users in age group 18-29); judging from popular subreddits and interest groups, many are interested in technology and science.

1.2. The German Reddit Sphere

Just as subreddits vary widely in topic and contents, the linguistic phenomena associated with particular subreddits are highly diverse. While the language in some subreddits is relatively close to the standard, others have unique memes and practices; making them difficult for outsiders to understand. In some German subreddits, for instance, emoticons may be replaced by *Umlaut* characters:

:) → Ü :o → Ö :< → Ä

On a lexical and phraseological level, typical expressions commonly associated with online communication as well as Reddit-specific memes are often translated word for word or even morpheme for morpheme, leading to a humorous effect: *pfostieren* 'to post' (*Pfosten* is a post in the sense of a pillar; the word is not usually associated with submitting online content), *ausgelöst* 'triggered', *Unterlases* 'subreddit', *fixierte das für dich* 'fixed this for you', *kantig* 'edgy' (in the sense of trying too hard to be avant-garde), *Kantenfürst* 'edgelord'.

1.3. The GeRedE Corpus

Currently, the domain of computer-mediated communication (CMC) is under-represented in the German corpus landscape. GeRedE aims to contribute to filling this gap. Due to the diversity of linguistic registers, we deem Reddit data a particularly valuable addition to corpus resources. Besides being linguistically interesting, Reddit can be – just like Twitter (Mejova et al., 2015) – regarded as a "digital socioscope" in the sense that it provides a window into various aspects of societal interaction.

[1] Ranks as of November 30, 2019: Alexa: 6 (US), 9 (DE); SimilarWeb: 11 (US), 33 (DE).
https://www.alexa.com/topsites/countries
https://www.similarweb.com/top-websites

[2] https://www.redditinc.com

[3] https://www.pewresearch.org/internet/fact-sheet/social-media

[4] https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020) , pages 6312–6318
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

This is exemplified by the case studies that have been conducted using various strands of discourse and argumentation studies, network analyses or knowledge transfer (see section 2.). In contrast to Twitter, Reddit presumably allows for such phenomena to be studied in greater depth. We assume that Reddit conversations are possibly more in-depth, given the different conditions of the platforms: while Twitter only allows for 280 characters per Tweet and reply, comment length on Reddit is limited to 10,000 characters, and submissions can include up to 40,000 characters.[5] Moreover, the data is freely available; making it a viable choice from a pragmatic point of view (cf. section 3.1.). The ever-growing popularity of the platform yields a vast amount of data; which is highly beneficial to increasing the reliability of quantitative corpus studies. Over time, the data base for GeRedE can likely be extended even further – this should be comparatively easy with the processing pipeline in place.

One of our main goals is not to provide an opportunistic sample of posts, but rather a collection that is as complete as possible, containing full conversation threads.

2. Related Work

Although several case-studies have been carried out on Reddit so far, the data collection has – to the best of our knowledge – been mostly ad-hoc. At the same time, they prove Reddit's potential for a large variety of research questions.

Mueller (2016) studies strawman argumentation on a number of hand-picked posts containing '<s>' to denote sarcasm. He shows that particular subreddits tend to have rather homogeneous user-bases; making pseudo-arguments a popular rhetorical strategy to construct and simultaneously discredit a hypothetical opponent for the purpose of community-building. This analysis is in line with Mitchell and Lim (2018), who argue that despite being accessible to everyone, Reddit's content is not egalitarian in that the moderators – a small number of very active users – decide which content is appropriate – an effect potentially enforced by the ranking algorithms focusing on up and downvote ratios. In a similar vein, the network studies by Buntain and Golbeck (2014) and Olson and Neal (2015) suggest that users tend to be very focused on one particular community and that it is very unlikely for somebody to regularly post on different subreddits. From other social networks, it is known that a vocal minority can dominate the discourse. For example, on Twitter, only 10% of American adult users of Twitter account for 80% of all tweets from this group.[6]

Working on a sample of submissions from different *Ask* subreddits, Kumar et al. (2018) demonstrate Reddit's usefulness for (the study of) knowledge transfer for various topics. They develop a codebook and show that subreddits show very different preferences for interaction types. For instance, "Socialising with negative intent" is common on *r/ask_politics*, but almost absent from the other subreddits considered (Kumar et al., 2018, 1940).

Tsou (2016) conducts a quantitative study on highly popular subreddits and uses ANOVA to determine their variability in terms of GIF and image use, sentiment and readability. Again, large disparities are identified between subreddits: the readability of *r/philosophy* is considerably lower than for *r/gifs*, and subreddits classified as *health/ food* have a much higher rate of emotional expressions than is the case for those classified as *news*.

3. Corpus Creation

3.1. Data Basis

In an ongoing effort, Jason Baumgartner has been collecting every Reddit submission and comment since 2005 and made them publicly accessible via `https://files.pushshift.io/reddit/`[7] (some caveats apply, see Gaffney and Matias (2018)). For both submissions and comments, compressed line-delimited JSON files are available for download, containing on a monthly basis all content created as well as the associated metadata. As Reddit has grown substantially over the years, so too have these files – although heavily compressed, a month's worth of data from 2019 amounts to over 15 GB. While an official API to access Reddit data also exists[8], the *Pushshift* files also include comments, submissions and subreddits that have since been deleted or banned from the site.

The first attempt at extracting German comments in particular 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). However, activity on Reddit has greatly increased since then, and German communities have thrived. Figure 1 shows the monthly comments in *r/de*, one of the oldest German subreddits. It is obvious that the amount of data created before 2015 is negligible compared to the amount of data created since then. This overall development can also be seen in Figure 2, showing the next four biggest German subreddits. Here, the volatile nature of Reddit is clearly visible: Subreddits can quickly rise to prominence, only to fall out of favour soon after – or to be permanently banned, as in the case of *r/edefreiheit*, a cesspool of the German alt-right where anti-refugee and anti-islamic comments abound.

3.2. Detecting German Posts

To detect German comments in the vast dataset, we first applied Barbaresi's approach of a two-tiered filter based on spell-checking and character-based language identification (Barbaresi, 2015). After some sanitizing steps to exclude extremely short (or deleted) comments, every token is checked against a German and an English dictionary. A comment is classified as potentially German if at least 70% of its tokens are found in the German 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

[5] Before June, 2015, this applied just to subreddits where only text submissions were allowed; other subreddits had a character limit of 15,000 for submissions.

[6] `https://www.pewresearch.org/fact-tank/2019/08/02/10-facts-about-americans-and-twitter/`

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

[8] See `https://www.reddit.com/dev/api`

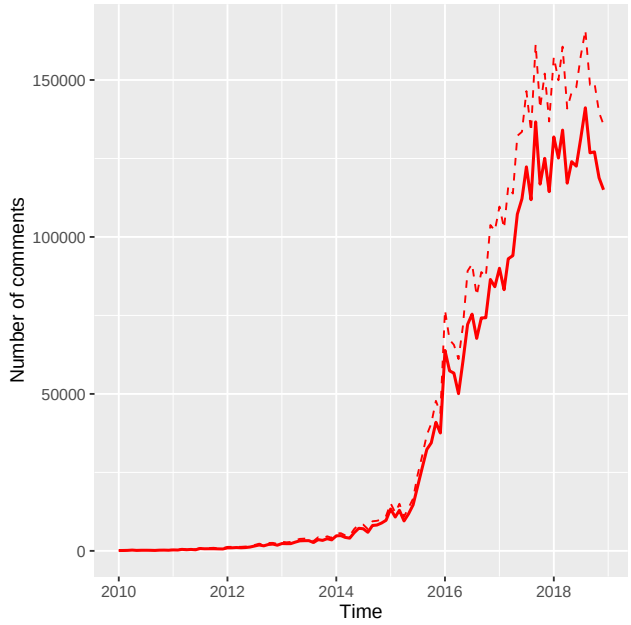

Figure 1: Comments per month classified as German in *r/de*, 2010-2018 (dashed line represents all comments)

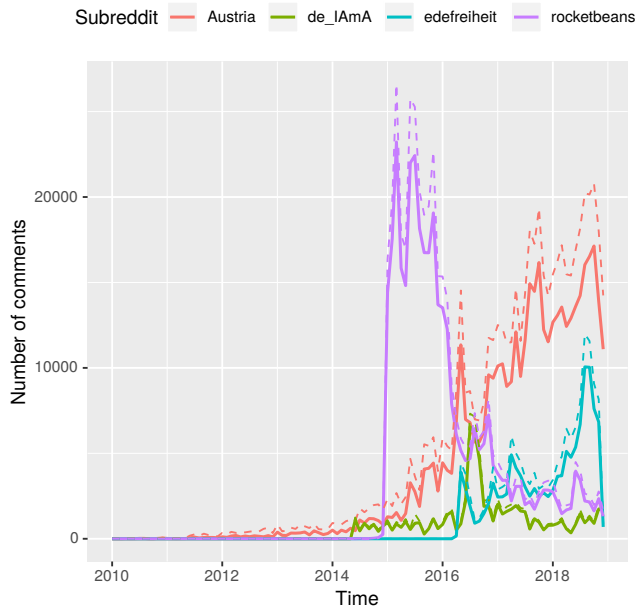

Figure 2: Subreddits with the most comments classified as German (after *r/de*), 2010-2018 (dashed line represents all comments)

ultimately classify comments as German. This way, we identified 8,753,873 comments between 2010 and 2018, almost half of them in *r/de*.

In a pre-study (Blombach et al., 2019), we evaluated the two-tiered filter's performance. We manually checked 565 randomly selected comments from 2016 and found that 148 (26%) of them were not actually in German. A closer inspection revealed that while longer comments are recognized rather reliably, shorter comments are often misclassified

Subreddit	"German"	Total	Fraction
de	4,050,574	4,839,414	0.837
Austria	434,218	557,630	0.779
rocketbeans	357,840	410,390	0.872
AskReddit	260,396	393,689,540	**0.000661**
edefreiheit	121,370	147,150	0.825
Fireteams	75,488	4,990,298	**0.0151**
de_IAmA	72,272	82,041	0.881
funny	68,898	78,100,018	**0.000882**
soccer	56,913	37,661,734	**0.00151**
pics	51,643	71,697,706	**0.000720**

Table 1: Comment counts in the 10 subreddits containing the highest number of comments classified as German with Barbaresi's approach (2010–2018).

for instance, due to the presence of proper names[9] and tokenization problems.

Table 1 shows the top 10 subreddits containing comments identified as German. A very low relative frequency of German comments in a subreddit indicates false positives which would negatively affect the precision of the two-tiered filter. At the same time, the fraction of German comments is usually below 90% – even for larger subreddits which are, to the best of our knowledge, almost exclusively in German. We interpret this as an indication that recall could also be improved. The reason for this recall problem is probably that the filter excludes extremely short comments. While these comments might not be too interesting in themselves, we would like to retain them to keep conversation threads intact.

We therefore improve upon the filter by implementing two additional steps. The first step is to ignore all subreddits containing too small fractions of German posts relative to the total number of comments in the subreddit. For subreddits with a high total number of comments, we only keep those with at least 1.5% of German comments (e. g., *r/germany* with 2.6%, *r/Switzerland* with 2.8% or *r/fcbayern* with 1.8%). For subreddits with a low total number of comments, this threshold has to be higher, since it would not make much sense to retain a subreddit with a total of say 10 comments of which one has been classified as German.

After testing several manually defined rough thresholds for different subreddit sizes, we used a function that closely matched the best thresholds to get a smoother filter curve:

$$e^{-\frac{\sqrt{N}}{4}} + 0.015, \quad (1)$$

where N is the number of all comments in a subreddit. Filtering out all subreddits with a fraction of 'German' comments below this dynamic threshold and removing 6,009 comments with duplicate IDs leaves us with 2,429 subreddits out of 43,193, containing 5,932,308 'German' comments. In other words, we filtered out approx. 32.2% of the comments in this step.

Since for many discourse analytic questions researchers will be interested in the actual conversation threads, we require

[9] For example, the filter often classifies comments mentioning German or German-sounding football players as German; the same goes for comments mentioning places in Germany, Austria or Switzerland.

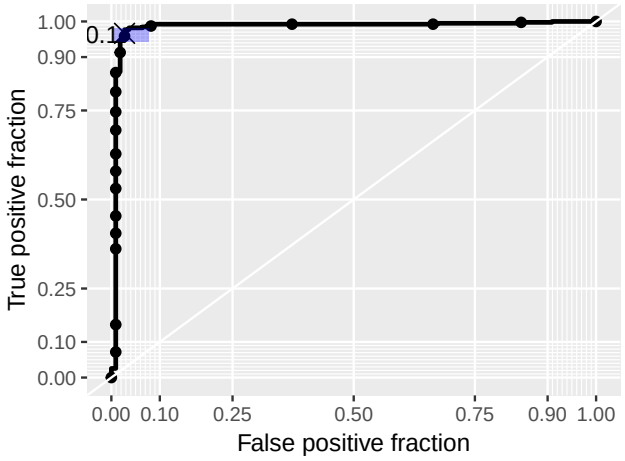

Figure 3: Receiver operating characteristic (ROC) curve for the thread filter based on a manually annotated, stratified sample of 500 threads

a procedure that yields complete threads. We thus reconstructed the threads by combining comments belonging to the same submission and took a stratified sample of 500 threads.[10] This sample was then manually annotated. Fortunately, most threads are either completely German or not German at all: Out of the 500 selected threads, 379 were indubitably German and 98 not German at all (amounting to more than 95%). 11 threads contained a significant amount of German (usually alongside English), and 12 threads contained *some* German words and sentences but were mostly English.[11]

We opted for a language-filtering approach combining both the number of comments identified as German in the thread at hand (N_{german}) and the relative number of German comments in the whole subreddit that the thread belongs to ($p_{\text{subreddit}}$). Let N_{thread} denote the number of comments in a given thread, then

$$\text{score}_{\text{thread}} = p_{\text{subreddit}} \cdot \frac{N_{\text{german}}}{N_{\text{thread}}}. \quad (2)$$

Figure 3 shows the receiver operating characteristic curve of our approach: With a cut-off of 0.1 for $\text{score}_{\text{thread}}$, we can give a 95% confidence interval for sensitivity (true positive rate) of $[95\%, 98\%]$ and of $[1\%, 8\%]$ for false positive rate. We are thus not far away from perfect classification at $(0, 1)$.

Filter	Comments	Threads	Subreddits
Two-tiered	8,753,873	–	43,193
Subreddits	5,932,308	511,996	2,429
Threads	6,786,591	388,680	1,977

Table 2: Number of comments after each filtering step

Table 2 shows the results of successively applying the filters described above. While our final filter further reduces the number of threads and subreddits in the corpus, we get more comments than after the second step, since we include all comments from included threads, with some not initially classified as German. Our final corpus contains comments by 141,409 different users (plus those whose author has since been deleted). The total number of tokens from submissions and comments is 271,679,444. Table 3 shows statistics for the 20 subreddits with the most tokens. Some differences between them can be clearly seen. The German version of "Ask me anything" (*r/de_IAmA*), for example, has many more comments per thread than most subreddits, indicating the higher importance of individual threads. The number of tokens per comment, on the other hand, may indicate how deep discussions typically go in a particular subreddit.[12]

```
<s>
So          ADV     ADV     DM
bin         VAFIN   VAFIN   VAFIN
endlich     ADV     ADJD    ADV
zu          APPR    APPR    APPR
was         PIS     PIS     PIS
gekommen    ADJD    VVPP    VVPP
:D          ADJD    EMOASC  EMOASC
</s>
<s>
Habe        VAFIN   VAFIN   VAFIN
...
```

Figure 4: Tagging and tokenization example as verticalized text. Columns are word form, POS-tag by Treetagger, POS-tag by SoMeWeTa, and manual correction of the POS tag.

4. Annotation

4.1. Pre-processing

Reddit comments can include markup in a Reddit-specific Markdown flavor. Luckily, Reddit's own snudown parser[13] is freely available and can be used to convert the Markdown to HTML.

4.2. Tokenization and Part-of-Speech Tagging

Aside from the Markdown markup, some of Reddit's peculiarities pose challenges to existing tools. Short form links to subreddits start with */r/* or *r/* (*r/wasletztepreis*), those to user profiles with */u/* or *u/* (*u/username*), which has to be accounted for by a tokenizer. Punctuation is often omitted and replaced by line breaks or emoticons (cf. Figure 4).

We tokenize the Reddit comments using SoMaJo (Proisl and Uhrig, 2016), to which we added support for Reddit short links. SoMaJo is able to make use of the information in the HTML output of the previous step to improve sentence boundary detection. The tokenized text is tagged with the STTS_IBK tagset (Beißwenger et al., 2015) using SoMeWeTa (Proisl, 2018). A post-processing script ensures that Reddit-specific phenomena that could be tagged deterministically but are unknown to the tagger (e. g. use of

[10] We stratified according to the score in equation 2.

[11] For the two-way classification that Figure 3 is based on, we categorized threads that contained a significant amount of German as German and the ones with only the occasional German sentence as non-German.

[12] In the case of *r/afdwatch*, a subreddit dedicated to monitoring the German far-right party AfD, threads mostly consist of links to newspaper articles and long quotations, explaining the high number of tokens per comment.

[13] `https://github.com/reddit/snudown/`

Subreddit	Comments	Tokens	Comments per thread			Tokens per comment		
			Mean	Median	SD	Mean	Median	SD
de	4799226	182430443	26.9777061	10	56.8233799	36.9003619	18	62.6738817
Austria	510669	20898949	18.8181818	10	26.1851997	39.1626455	20	63.6346196
rocketbeans	408090	17975825	17.2102733	8	32.3833304	39.7930187	22	61.3329569
edefreiheit	144085	6749444	7.2306418	4	12.0093476	43.5746191	20	86.5951918
de_IAmA	81480	5291808	40.6181456	21	57.5620299	62.9244845	32	95.9035496
Finanzen	53309	3675139	17.5358553	12	18.7072148	61.5715733	38	76.3257501
FragReddit	35519	1767716	13.5568702	10	12.1702089	43.6782849	23	64.6584417
PietSmiet	26628	1643558	8.9445751	6	10.5857967	51.6566021	27	80.9625971
wien	39395	1520977	9.6296749	7	9.9211557	34.0579007	19	50.9778974
afdwatch	6780	1296656	3.0389960	2	4.1096849	187.2781711	152	167.8638880
German	13591	1264269	5.9740659	4	7.3829021	69.7810316	27	132.4196588
rbtvcirclejerk	38865	1159325	17.4988744	13	16.6244860	27.4095201	15	44.6759074
einfach_posten	24032	1089008	8.2527473	6	7.0068421	34.0825566	17	57.1009909
kreiswichs	54034	1007592	6.8859437	5	6.6144594	14.3191324	7	47.1998514
Weibsvolk	12172	923521	11.8174757	8	18.4001438	71.3578705	40	103.6198935
DSA_RPG	8876	859029	10.0180587	8	9.0191078	82.8406940	48	107.5028048
de_EDV	14184	834322	9.9119497	7	10.1982072	48.5414552	30	66.5873373
Dachschaden	11099	803411	8.3892668	4	14.8284830	65.4934679	28	118.0158623
de_simulator	21522	776608	10.8696970	10	6.3110538	34.6621132	20	49.7357932
the_schulz	32448	742743	7.1692444	4	15.8072671	20.0840422	11	34.5227534
deutschland	14859	724392	6.9305037	4	8.9442887	44.5753415	22	70.9594945

Table 3: Statistics for the top 20 subreddits in the final corpus (note that the number of tokens of the submissions themselves is also included in the calculation of the total number of tokens under "Tokens")

Umlaut characters as emoticons or short links) are tagged correctly.

4.3. Evaluation of Part-of-Speech Tagging

We manually corrected the part-of-speech tags in a random sample of comments (1,150 tokens) to evaluate performance. The tagging accuracy was at 92.35% (95% Wilson score interval: 90.67–93.75%). This falls between the values reported for SoMeWeTa's performance on CMC (89.06%) and web data (93.75%) and indicates that Reddit posts deviate less from standard German than other CMC varieties. POS errors seem 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, some 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**).

4.4. Lemmatization

Deviations from the norms of the written standard language can make the lemmatization of CMC data a challenging task. We evaluated different baseline strategies and off-the-shelf lemmatizers against a corpus of web and CMC texts (Proisl et al., 2020) and found that the best-performing tool was a thin wrapper around the SMOR finite-state morphological analyzer (Schmid, 2004). With a bit of post-processing and some additional heuristics, SMOR achieved an approximately human-level accuracy of 96.20%.[14] To lemmatize GeRedE, we followed the same procedure.

5. Data Access and Metadata

We provide the corpus in a version that was indexed with the IMS Open Corpus Workbench (Evert and Hardie, 2011). After registration, it can be accessed by academic users via CQPweb (Hardie, 2012) at `https://corpora.linguistik.uni-erlangen.de/cqpweb/gerede_v1/`.

The following thread-level metadata are available as categorical variables in CQPweb, meaning they can be used to create subcorpora:

- "text_year": the year the submission was posted to Reddit
- "text_subreddit_class": subreddit name (only the top 100 can be chosen here)
- "text_link_flair_text_class": subreddit-specific tag assigned to the submission, usually used to filter a subreddit for specific content, e. g. politics or humour (only the 100 most common tags can be chosen here); Table 4 shows statistics for the 20 tags with the most tokens (indicating, for example, that comments on humorous submissions tend to be much shorter)
- "text_over_18": whether the submission has been tagged as NSFW (not suitable for work, e. g. containing adult content)
- "text_gilded": whether the thread was gilded

[14] The wrapper script is available here: `https://github.com/fau-klue/empirist-corpus/tree/master/baselines`.

Flair	Comments	Tokens	Comments per thread			Tokens per comment		
			Mean	Median	SD	Mean	Median	SD
Frage/Diskussion	830051	34449809	39.6	19	63.6	38.8	19	64.3
Politik	606845	26412898	27.0	8	68.0	42.6	21	71.4
Nachrichten	418227	18999574	22.0	9	39.8	45.0	24	71.2
Humor	556721	13011245	25.8	10	46.6	22.9	12	40.8
Interessant	239413	8407645	25.2	9	52.2	34.3	17	59.3
Nachrichten DE	176406	7254230	28.8	9	52.8	40.7	22	63.8
Gesellschaft	154444	6977392	35.5	12	65.8	44.4	22	73.9
Frage	146802	6623095	18.0	11	23.7	40.6	22	60.5
Flüchtlinge	124176	6221093	30.7	12	50.6	49.4	26	79.0
Diskussion	102930	5628369	25.7	13	45.8	49.8	26	79.6
TIRADE	129313	5287980	68.5	29	103.	38.2	20	59.3
Medien	113127	4262804	23.9	9	43.5	36.9	19	60.9
Wissenschaft&Technik	100318	4232370	17.0	6	33.2	41.3	22	61.9
Kriminalität	103797	3895856	29.9	10	54.7	37.0	20	56.8
Feedback	55346	2909881	34.2	18	47.7	48.8	28	72.8
Humor/MaiMai	126607	2864402	32.7	11	70.5	22.3	11	38.9
Nachrichten Europa	63518	2560498	25.2	8	50.1	39.8	21	63.1
Nachrichten Welt	61015	2502916	20.3	7	40.3	40.4	21	64.6
Nachrichten Deutschland	61545	2405965	26.6	9	49.3	38.7	21	59.0
Wirtschaft	52891	2314995	19.3	7	43.1	43.1	23	65.9

Table 4: Statistics for the top 20 submission flair texts (note that the number of tokens of the submissions themselves is also included in the calculation of the total number of tokens under "Tokens")

- "text_distinguished": whether a submission or comment has been distinguished by a moderator or administrator

Additionally, users can use various types of thread- and comment-level metadata in their queries. As this data will be fully documented in CQPweb, we only include some of the more interesting metadata here:

- "ymd", "ym": exact day or month and year the submission was posted
- "subreddit": specific subreddit in which a submission or comment was posted
- "score": difference of upvotes and downvotes the submission or comment received[15]
- "author": name of the submission's or comment's author

Concerning the actual text of submissions and comments, paragraphs, sentences and various Markdown markup are also searchable.
For those who wish to recreate our corpus, we provide our whole set of scripts and annotation at `https://github.com/fau-klue/german-reddit-korpus`.

6. Outlook

Due to their peculiarities, Reddit data are a promising source for further (socio-)linguistic research. Since those very features pose challenges to existing tools, resources for tokenization and tagging may need to be updated.

In the future, we plan to include additional data from 2019 and Reddit's early years (2005–2009).[16] We also plan to provide a new annotation layer using an adjusted version of the STTS_IBK tagset. Furthermore, there is a lot of potential for additional metadata. For example, the current data already includes user flair texts, subreddit-specific short texts displayed next to a user's name. These flairs can either be set by the users themselves or be assigned by moderators. Since flairs often provide personal information like the user's residence or occupation, their content could be matched against lists of countries, federal states or cities. Although Reddit does not provide geotagging, information like this could be used to infer users' provenance which might be valuable e. g. for studies concerned with diatopical variation. On the user-level, we would also like to add information about users' activity levels, classifying e. g. influential power users (cf. section 2.).
Certain subreddits containing heavily dialectal language use (such as *r/aeiou* or *r/BUENZLI*) probably deserve special treatment. Comments from these subreddits are classified as German less reliably, and tokenization and POS tagging often fail spectacularly. We may therefore decide to release these data as a separate corpus in the future. The same applies to subreddits like *r/German* or *r/GermanPractice* frequented by learners of German (and native speakers who help them) – this data differs markedly from "regular" German subreddits, but might be of special interest to researchers studying second language acquisition.

7. Bibliographical References

Barbaresi, A. (2015). Collection, description, and visualization of the German Reddit corpus. In *Proceedings of the 2nd Workshop on Natural Language Processing*

[15] Note that scores may not reflect the true popularity of submissions and comments as the time of access to these data also plays a role – posts that were older when they were processed may have higher scores than newer ones.

[16] However, we do not expect a lot of posts from the latter period, given that we have not found many German posts from 2010.

for Computer-Mediated Communication / Social Media, pages 7–11, Essen.

Beißwenger, M., Bartz, T., Storrer, A., and Westpfahl, S. (2015). Tagset und Richtlinie für das Part-of-Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation. Guideline document.

Beißwenger, M., Bartsch, S., Evert, S., and Würzner, K.-M. (2016). EmpiriST 2015: A shared task on the automatic linguistic annotation of computer-mediated communication and web corpora. In Paul Cook, et al., editors, *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 44–56, Berlin. Association for Computational Linguistics.

Blombach, A., Dykes, N., Evert, S., Heinrich, P., Kabashi, B., and Proisl, T. (2019). A new German Reddit corpus. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Kaleidoscope Abstracts*, pages 278–279, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Buntain, C. and Golbeck, J. (2014). Identifying social roles in reddit using network structure. In *Proceedings of the 23rd international conference on world wide web*, pages 615–620. ACM.

Evert, S. and Hardie, A. (2011). Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*, University of Birmingham, UK.

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

Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

Kumar, P., Gruzd, A., Haythornthwaite, C., Gilbert, S., Esteve del Valle, M., and Paulin, D. (2018). Learning in the wild: Coding reddit for learning and practice. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.

Lui, M. and Baldwin, T. (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.

Yelena Mejova, et al., editors. (2015). *Twitter: A Digital Socioscope*. Cambridge University Press, New York.

Mitchell, S. S. and Lim, M. (2018). Too crowded for crowdsourced journalism: Reddit, portability, and citizen participation in the Syrian crisis. *Canadian Journal of Communication*, 43(3):399–419.

Mueller, C. (2016). Positive feedback loops: Sarcasm and the pseudo-argument in Reddit communities. *Working Papers in TESOL & Applied Linguistics*, 16(2):84–97.

Olson, R. S. and Neal, Z. P. (2015). Navigating the massive world of reddit: Using backbone networks to map user interests in social media. *PeerJ Computer Science*, 1:e4.

Proisl, T. and Uhrig, P. (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.

Proisl, T., Dykes, N., Heinrich, P., Kabashi, B., Blombach, A., and Evert, S. (2020). Empirist corpus 2.0: Adding manual normalization, lemmatization and semantic tagging to a German web and CMC corpus. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille. ELRA.

Proisl, T. (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.

Schmid, H. (2004). Efficient parsing of highly ambiguous context-free grammars with bit vectors. In *Proceedings of COLING 2004*, Geneva, Switzerland.

Tsou, A. (2016). How does the front page of the internet behave? Readability, emoticon use, and links on Reddit. *First Monday*, 21(11).

8. Language Resource References

Michael Beißwenger and Sabine Bartsch and Stefan Evert and Kay-Michael Würzner. (2016). *EmpiriST 2015 Gold Standard*. `https://sites.google.com/site/empirist2015/home/gold/`.

Sabine Brants and Stefanie Dipper and Peter Eisenberg and Silvia Hansen and Esther König and Wolfgang Lezius and Christian Rohrer and George Smith and Hans Uszkoreit. (2004). *TIGER Corpus*. IMS Stuttgart, `http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html`.

Thomas Proisl and Peter Uhrig. (2016). *SoMaJo*. `https://github.com/tsproisl/SoMaJo`.

Thomas Proisl. (2017). *SoMeWeTa*. `https://github.com/tsproisl/SoMeWeTa`.

Helmut Schmid. (2002). *SMOR*. `https://www.cis.uni-muenchen.de/~schmid/tools/SMOR/`.