A New German Reddit Corpus

A Report on Work in Progress

Andreas Blombach, Natalie Dykes,
Philipp Heinrich, Thomas Proisl

Chair of Computational Corpus Linguistics
Friedrich-Alexander-Universität Erlangen-Nürnberg

October 10, 2019

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

What's Reddit?

- Social news aggregation, discussion and micro-blogging platform
- Founded in 2005
- One of the most popular websites in the US (Alexa rank: 6[1], SimilarWeb rank: 10[2]
- Increasingly popular in Germany (Alexa rank: 9[3], SimilarWeb rank: 31[4]

[1] https://www.alexa.com/topsites/countries/US, 9.10.2019
[2] https://www.similarweb.com/top-websites/united-states, 9.10.2019
[3] https://www.alexa.com/topsites/countries/DE, 9.10.2019
[4] https://www.similarweb.com/top-websites/germany, 9.10.2019

What's Reddit?

- Self-proclaimed "front page of the internet":
  - Users submit content (e.g. links, images, text)
  - Submissions are voted up or down
  - Submissions with the most upvotes make it to the front page
- Submissions can be commented on, comments can also be voted up or down
- Nested conversation threading
- Since 2008, users ("redditors") can create "subreddits" (categories or communities) with their own rules

Who uses Reddit?

- While more women than men use social media in general (at least in the US[5], Reddit is more popular among men than among women[6]
- Users are mostly young[7], with a high share of people interested in technology

[5] https://www.statista.com/statistics/471345/us-adults-who-use-social-networks-gender/

[6] https://www.statista.com/statistics/261765/share-of-us-internet-users-who-use-reddit-by-gender/

[7] https://www.statista.com/statistics/261766/share-of-us-internet-users-who-use-reddit-by-age-group/

The German Reddit Sphere

- Subreddits can be very heterogenous
- Parts of Reddit can be very weird to outsiders, the same goes for large parts of the German Reddit sphere
- Special emoticons: :) $\rightarrow$ Ü, :o $\rightarrow$ Ö, :< $\rightarrow$ Ä
- Quite common: word-for-word translations of English words and phrases:
  - *pfostieren*
  - *ausgelöst*
  - *Maimai*
  - *Fixierte das für dich.*
  - *schuldiges Vergnügen*
  - *Lases*, *Unterlases*
  - *kantig*, *Kantenfürst*

Building a Reddit Corpus

- In an ongoing effort, Jason Baumgartner collects every Reddit submission and comment, publicly accessible via `https://files.pushshift.io/reddit/`[8] (some caveats apply, see Gaffney and Matias, 2018)
- 2015: first attempt at extracting German comments (Barbaresi, 2015), resulting corpus still relatively small (97505 comments, 566362 tokens)
- Since then, activity on Reddit has increased greatly
- We adapt Barbaresi's approach of using a two-tiered filter to detect German comments in the vast dataset:
  1. spell checking
  2. language identification

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

Raw data (1)


```
'archived': False,
'author': 'Mythic_Emperor',
'author_created_utc': 1521765799,
'author_flair_background_color': None,
'author_flair_css_class': None,
'author_flair_richtext': [],
'author_flair_template_id': None,
'author_flair_text': None,
'author_flair_text_color': None,
'author_flair_type': 'text',
'author_fullname': 't2_12w6sr28',
'author_patreon_flair': False,
'body': 'Es geht mir Gut. Und dir?',
'can_gild': True,
'can_mod_post': False,
'collapsed': False,
'collapsed_reason': None,
'controversiality': 0,
'created_utc': 1541030413,
```

Raw data (2)


```
'distinguished': None,
'edited': False,
'gilded': 0,
'gildings': {'gid_1': 0, 'gid_2': 0, 'gid_3': 0},
'id': 'e8tkilo',
'is_submitter': False,
'link_id': 't3_9t23an',
'no_follow': True,
'parent_id': 't1_e8tfeac',
'permalink': '/r/HistoryMemes/comments/9t23an/are_you_just_gonna_scroll
'removal_reason': None,
'retrieved_on': 1544848339,
'score': 2,
'send_replies': True,
'stickied': False,
'subreddit': 'HistoryMemes',
'subreddit_id': 't5_2v2cd',
'subreddit_name_prefixed': 'r/HistoryMemes',
'subreddit_type': 'public'
```

Detecting German posts – the strategy

Barbaresi (2015)

1. Sanitizing
   - Ignore newlines and URLs
   - Don't consider very short posts or posts which were deleted
2. Spell checking
   - Check every token (`re.findall(\w+)`) against both a German and an English dictionary
   - If not in respective dictionary: count an error
   - Classify as potentially German if English error rate higher than 30% and German error rate less than 70%
3. Language identification
   - Run langid (Lui and Baldwin, 2014) on potentially German posts

Evaluation of the Two-tiered Filter

- Random sample of 1618 comments from 2016
- Looking at the first 565 comments (20020 lines), we see many of them (148, or 26%) are not actually in German (1423 lines $\rightarrow$ 7%)
- Longer comments are apparently recognized quite reliably as German, while shorter ones are often misclassified
- Also problematic: comments which only contain links to subreddits, lots of markup or proper names

Possible Corpus Cleanup

- Top 10 subreddits, 2016-2018 (out of 33144, containing 6784837 "German" comments):

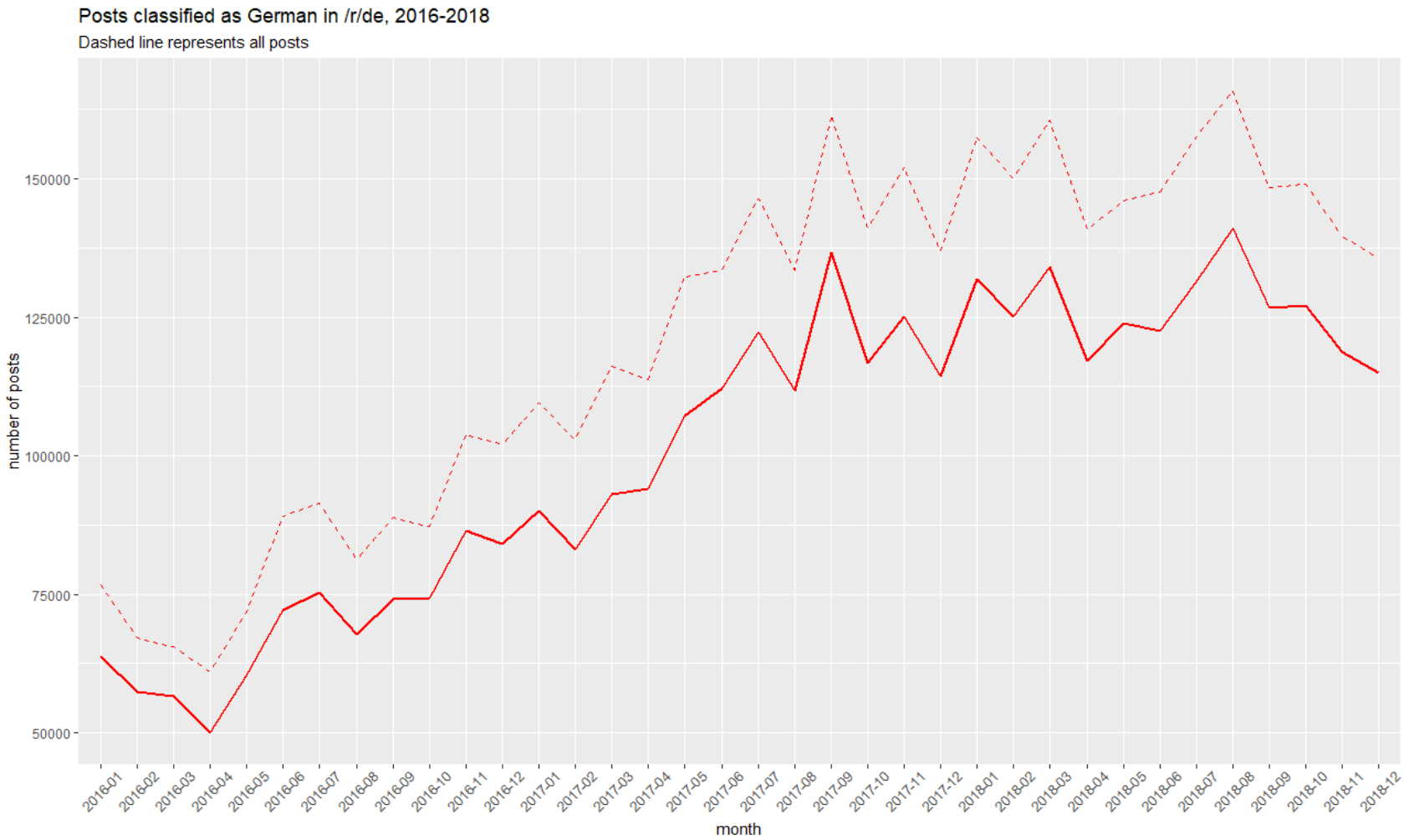
Subreddit	"German"	Total	Fraction
de	3644481	4364440	0.835
Austria	389041	483125	0.805
rocketbeans	142787	164699	0.867
AskReddit	125411	183147454	0.000685
edefreiheit	121370	147150	0.825
de_IAmA	58172	65768	0.885
Finanzen	48033	55009	0.873
Fireteams	38042	2125695	0.0179
soccer	33727	22918204	0.00147
rbtvcirclejerk	32493	38945	0.834

- Comments in subreddits with low fractions of "German" comments are probably not German after all

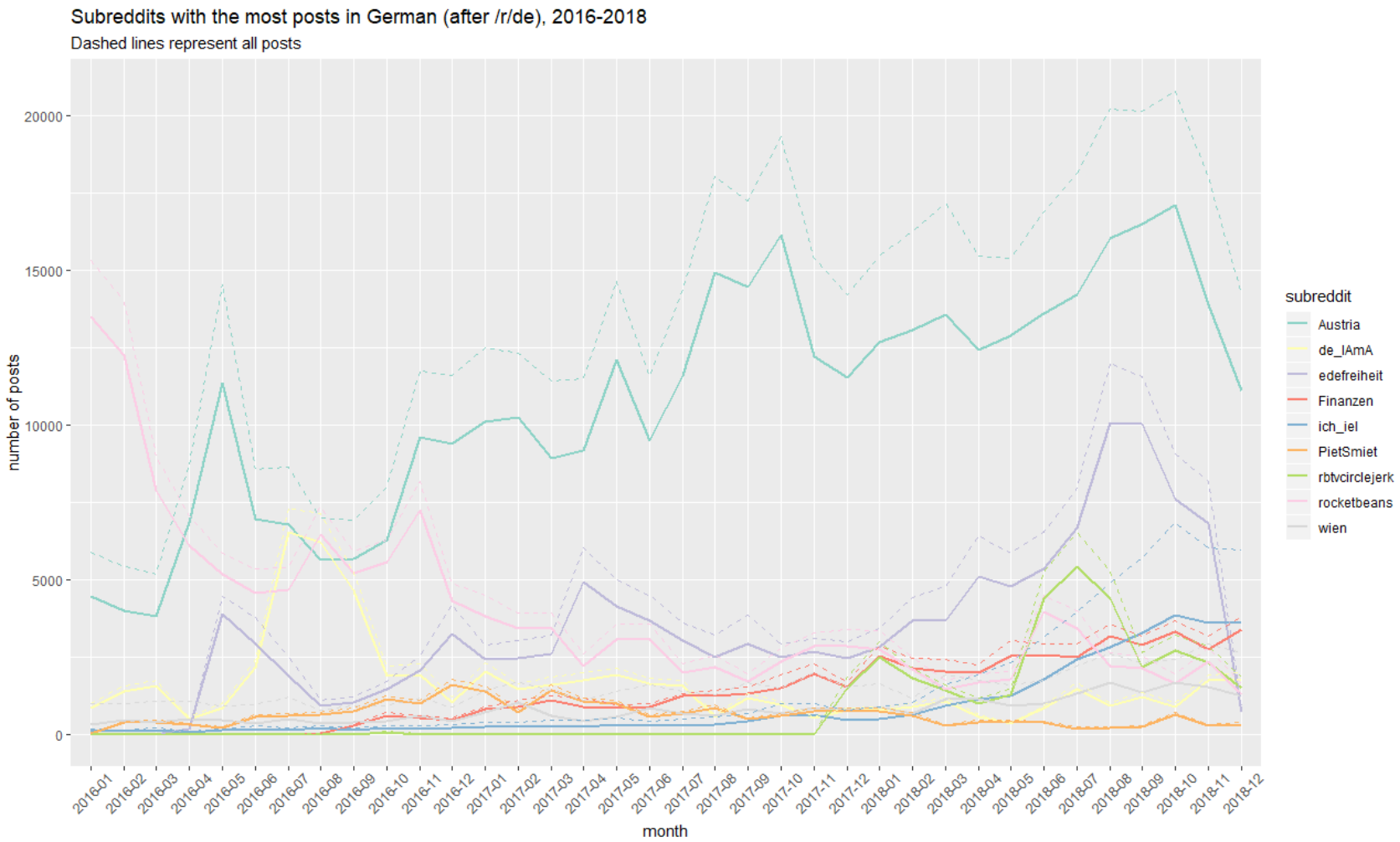
Possible Corpus Cleanup

- By filtering out subreddits with less than 10% and/or less than ten "German" comments, we can remove 27% of all comments classified as German
- Remaining comments: 4926924
- However, for those subreddits which are predominantly German, shouldn't we include *all* comments in our corpus?

Top 10 Subreddits



Top 10 Subreddits



Pre-processing and Tokenization

- Reddit comments can contain Markdown markup[9], so they have to be pre-processed to avoid curious tokenization results:
  - *∼∼der gute∼∼ Ketchup gehört in den ∼∼Reis oder aufs Brot∼∼ Müll.*
  - *Ich würde dir die [∼∼Sokratische Methode∼∼](https://de.wikipedia.org/wiki/Sokratische_Methode) [Mäeutik](https://de.wikipedia.org/wiki/M%C3%A4eutik) empfehlen.*
- Short form links to subreddits and user pages have to be tokenized properly (*r/oldschoolcool*, */r/de*, */r/de/about/traffic*, *l/de*, */u/username*, */u/username*, ...)
- In the absence of punctuation, line breaks (and sometimes, emoticons) can mark sentence boundaries

[9] see `https://www.reddit.com/wiki/markdown`

Example: Emoticons as Sentence Boundaries

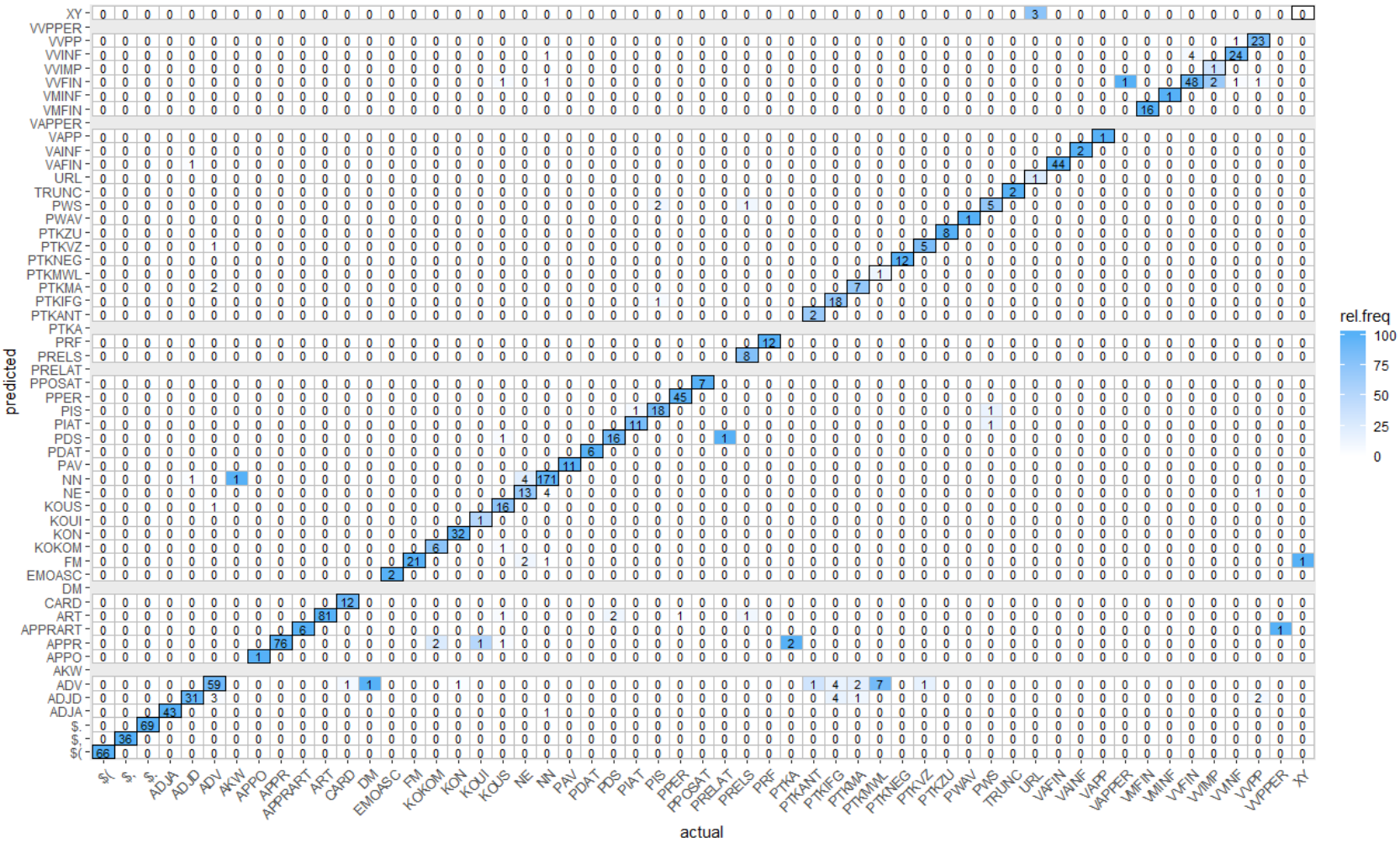
Token	TreeTagger	Lemma	SoMeWeTa	Corrected POS
<s>				
So	ADV	so	ADV	DM
bin	VAFIN	sein	VAFIN	VAFIN
endlich	ADV	endlich	ADJD	ADV
zu	APPR	zu	APPR	APPR
was	PIS	was	PIS	PIS
gekommen	ADJD	kommen	VVPP	VVPP
:D	ADJD	:D	EMOASC	EMOASC
Habe	VAFIN	haben	VAFIN	VAFIN
jetzt	ADV	jetzt	ADV	ADV
...				

Tokenization and POS Tagging

- Random sample tokenized with SoMaJo[10] and tagged with SoMeWeTa[11] (trained on EmpiriST corpus; tagset: STTS_IBK[12])
- Corrected POS tags in a small number of comments (1186 tokens – so far) to evaluate performance
- Accuracy: 92.6%

[10] Proisl and Uhrig (2016)
[11] Proisl (2018)
[12] Beißwenger et al. (2015)

Confusion Matrix



Outlook

- More corpus cleaning necessary
- Tokenization rules have to be updated
- POS-tagging, lemmatization etc. further down the pipeline
- Tagset: more discussion needed

Bonus Content: Some Notes on STTS-IBK

- Some very fine-grained categories like DM (e.g. "epistemisches *weil*"), PTKIFG, PTKMWL which are difficult to annotate manually, but no differentiation between some more obvious ones, e.g. *sein* as an auxiliary or as a full verb, definite and indefinite articles, ...
- Tags for some contracted forms but not all (STTS-IBK guidelines state to just annotate POS for the first word in these cases)
- Catch-all category $(\rightarrow$ no differentiation between opening and closing brackets or quotation marks, dashes, slashes etc. (asterisks to mark "Aktionswörter" also fall into this category)
- What to do with acronyms like *scnr* (*sorry, could not resist*) or *imho* (*in my humble opinion*)? Do we need new tags?
- TRUNC nur für Kompositionserst-, nicht aber für -zweitglieder

References I

Adrien Barbaresi. Collection, Description, and Visualization of the German Reddit Corpus. In German Society for Computational Linguistics & Language Technology, editor, *2nd Workshop on Natural Language Processing for Computer-Mediated Communication*, Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media, pages 7–11, Essen, Germany, September 2015.

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

Devin Gaffney and J. Nathan Matias. Caveat emptor, computational social science: Large-scale missing data in a widely-published reddit corpus. *PLOS ONE*, 13(7): 1–13, 07 2018.

Marco Lui and Timothy Baldwin. Accurate language identification of twitter messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 17–25, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

References II

Thomas Proisl. SoMeWeTa: A part-of-speech tagger for German social media and web texts. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 665–670, Miyazaki, 2018. European Language Resources Association.

Thomas Proisl and Peter Uhrig. SoMaJo: State-of-the-art tokenization for German web and social media texts. In Paul Cook, Stefan Evert, Roland Schäfer, and Egon Stemle, editors, *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin, 2016. Association for Computational Linguistics.