NLP for German CMC Texts

Tokenization, POS Tagging, and a New Gold Standard for Lemmatization

Thomas Proisl, Natalie Dykes, Andreas Blombach, Philipp Heinrich, Besim Kabashi

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

September 17, 2019

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

1. Introduction
2. EmpiriST Corpus and Derived Tools
   - Tokenization
   - POS Tagging
3. Lemmatization
   - Guidelines
   - Annotation
   - Baseline systems
4. Ongoing project: Creation of a German Reddit corpus
   - Data Gathering
   - Corpus Creation
   - Annotation
5. Conclusion

EmpiriST 2015 shared task

Beißwenger et al. (2016)

- Computer-mediated communication (CMC) poses challenges to NLP tools:
  - `Eveeentuell suchte ich gerade das Spiel :D`
  - `Ein Distributor darf zB Verlag des gemeinsamen Besitzers auch nicht ggüber anderen Verlägen bevorzugen.`
  - `@Montaaag #semibk DieseFrage spiele ich zurück indieRunde: Zu welchemHandlungsbereich gehört unsereKomm hier? Bildung?Freizeit?Mischung?`
- Goal: Encourage adaptation of NLP tools to CMC/web texts
  - Tokenization
  - POS tagging
- Dataset: 22,858 tokens
  - CMC part: Tweets, chats, Wiki talk pages, blog comments
  - Web part: Websites, blogs, Wikipedia articles, Wikinews

EmpiriST 2015 shared task

Beißwenger et al. (2016)

- Computer-mediated communication (CMC) poses challenges to NLP tools:
  - `Eveeentuell suchte ich gerade das Spiel :D`
  - `Ein Distributor darf zB Verlag des gemeinsamen Besitzers auch nicht ggüber anderen Verlägen bevorzugen.`
  - `@Montaaag #semibk DieseFrage spiele ich zurück indieRunde: Zu welchemHandlungsbereich gehört unsereKomm hier? Bildung?Freizeit?Mischung?`
- Goal: Encourage adaptation of NLP tools to CMC/web texts
  - Tokenization
  - POS tagging
- Dataset: 22,858 tokens
  - CMC part: Tweets, chats, Wiki talk pages, blog comments
  - Web part: Websites, blogs, Wikipedia articles, Wikinews

EmpiriST 2015 shared task

Beißwenger et al. (2016)

- Computer-mediated communication (CMC) poses challenges to NLP tools:
  - `Eveeentuell suchte ich gerade das Spiel :D`
  - `Ein Distributor darf zB Verlag des gemeinsamen Besitzers auch nicht ggüber anderen Verlägen bevorzugen.`
  - `@Montaaag #semibk DieseFrage spiele ich zurück indieRunde: Zu welchemHandlungsbereich gehört unsereKomm hier? Bildung?Freizeit?Mischung?`
- Goal: Encourage adaptation of NLP tools to CMC/web texts
  - Tokenization
  - POS tagging
- Dataset: 22,858 tokens
  - CMC part: Tweets, chats, Wiki talk pages, blog comments
  - Web part: Websites, blogs, Wikipedia articles, Wikinews

EmpiriST tokenization guidelines in a nutshell

Beißwenger et al. (2015a)

- Usual treatment of whitespace and punctuation
- Whitespace errors usually not corrected (`schona ber`)
  - Except in emoticons and symbols:
    `: )` → `:)`, `f - > d` → `f -> d`
- Split CamelCase (`deineMutter`) if not proper name (`MySpace`)
- Abbreviations:
  - multidot abbreviations representing multiple words are split up:
    `d.h.` → `d. h.`, `u.dgl.` → `u. dgl.`
  - multidot abbreviations representing single words are not split up: `o.k.` → `o.k.`, `Dipl.-Ing.` → `Dipl.-Ing.`
- Contractions are not split up (↯ Universal Dependencies):
  `machste` → `machste`
- Single tokens:
  - HTML/XML tags, E-mail addresses, URLs, Filenames, Emoticons

EmpiriST tokenization guidelines in a nutshell

Beißwenger et al. (2015a)

- Usual treatment of whitespace and punctuation
- Whitespace errors usually not corrected (`schona ber`)
  - Except in emoticons and symbols:
    `: )` → `:)`, `f - > d` → `f -> d`
- Split CamelCase (`deineMutter`) if not proper name (`MySpace`)
- Abbreviations:
  - multidot abbreviations representing multiple words are split up:
    `d.h.` → `d. h.`, `u.dgl.` → `u. dgl.`
  - multidot abbreviations representing single words are not split up: `o.k.` → `o.k.`, `Dipl.-Ing.` → `Dipl.-Ing.`
- Contractions are not split up (↯ Universal Dependencies):
  `machste` → `machste`
- Single tokens:
  - HTML/XML tags, E-mail addresses, URLs, Filenames, Emoticons

EmpiriST tokenization guidelines in a nutshell

Beißwenger et al. (2015a)

- Usual treatment of whitespace and punctuation
- Whitespace errors usually not corrected (`schona ber`)
  - Except in emoticons and symbols:
    `: )` → `:)`, `f - > d` → `f -> d`
- Split CamelCase (`deineMutter`) if not proper name (`MySpace`)
- Abbreviations:
  - multidot abbreviations representing multiple words are split up:
    `d.h.` → `d. h.`, `u.dgl.` → `u. dgl.`
  - multidot abbreviations representing single words are not split up: `o.k.` → `o.k.`, `Dipl.-Ing.` → `Dipl.-Ing.`
- Contractions are not split up (↯ Universal Dependencies):
  `machste` → `machste`
- Single tokens:
  - HTML/XML tags, E-mail addresses, URLs, Filenames, Emoticons

EmpiriST tokenization guidelines in a nutshell

Beißwenger et al. (2015a)

- Usual treatment of whitespace and punctuation
- Whitespace errors usually not corrected (`schona ber`)
  - Except in emoticons and symbols:
    `: )` → `:)`, `f - > d` → `f -> d`
- Split CamelCase (`deineMutter`) if not proper name (`MySpace`)
- Abbreviations:
  - multidot abbreviations representing multiple words are split up:
    `d.h.` → `d. h.`, `u.dgl.` → `u. dgl.`
  - multidot abbreviations representing single words are not split up: `o.k.` → `o.k.`, `Dipl.-Ing.` → `Dipl.-Ing.`
- Contractions are not split up (↯ Universal Dependencies):
  `machste` → `machste`
- Single tokens:
  - HTML/XML tags, E-mail addresses, URLs, Filenames, Emoticons

EmpiriST tokenization guidelines in a nutshell

Beißwenger et al. (2015a)

- Usual treatment of whitespace and punctuation
- Whitespace errors usually not corrected (`schona ber`)
  - Except in emoticons and symbols:
    `: )` → `:)`, `f - > d` → `f -> d`
- Split CamelCase (`deineMutter`) if not proper name (`MySpace`)
- Abbreviations:
  - multidot abbreviations representing multiple words are split up:
    `d.h.` → `d. h.`, `u.dgl.` → `u. dgl.`
  - multidot abbreviations representing single words are not split up: `o.k.` → `o.k.`, `Dipl.-Ing.` → `Dipl.-Ing.`
- Contractions are not split up (↯ Universal Dependencies):
  `machste` → `machste`
- Single tokens:
  - HTML/XML tags, E-mail addresses, URLs, Filenames, Emoticons

EmpiriST tokenization guidelines in a nutshell

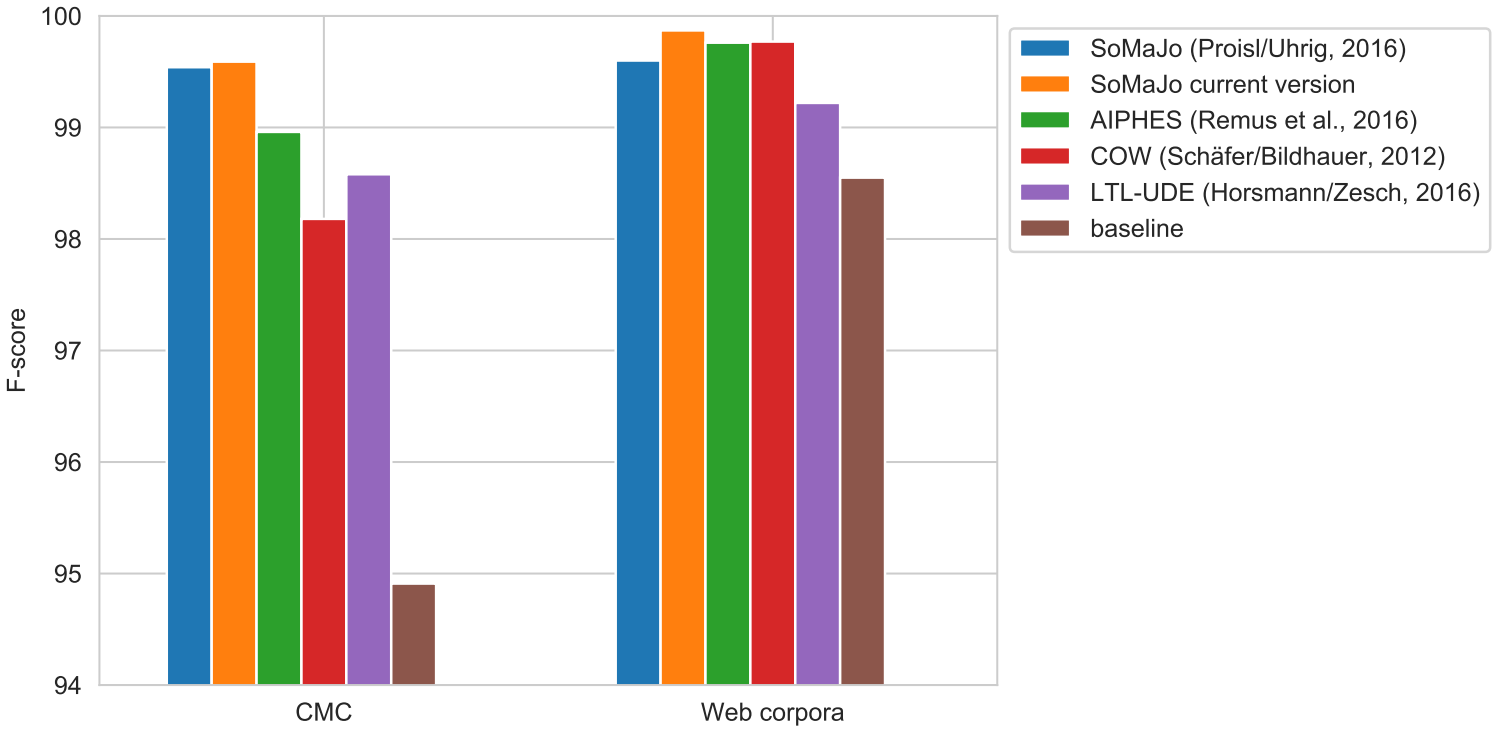
Beißwenger et al. (2015a)

- Usual treatment of whitespace and punctuation
- Whitespace errors usually not corrected (`schona ber`)
  - Except in emoticons and symbols:
    `: )` → `:)`, `f - > d` → `f -> d`
- Split CamelCase (`deineMutter`) if not proper name (`MySpace`)
- Abbreviations:
  - multidot abbreviations representing multiple words are split up:
    `d.h.` → `d. h.`, `u.dgl.` → `u. dgl.`
  - multidot abbreviations representing single words are not split up: `o.k.` → `o.k.`, `Dipl.-Ing.` → `Dipl.-Ing.`
- Contractions are not split up (↯ Universal Dependencies):
  `machste` → `machste`
- Single tokens:
  - HTML/XML tags, E-mail addresses, URLs, Filenames, Emoticons

Some of the problematic cases

- Abbreviations vs. actual words: `automat.`, `zum.`
- Cardinal number at end of sentence vs. ordinal number
- Section numbers vs. dates: `5.3.`
- Citations (`Storrer2007`) vs. proper names (`Blume2000`)
- Horizontal ellipsis:
  - `du bist echt ein Arm...` → `Arm...`
  - `zeig mir mal deinen Arm...` → `Arm ...`
- Confusion of hyphen (-) and en dash (–)

Results


100
99
98
97
96
95
94
F-score
CMC
Web corpora
SoMaJo (Proisl/Uhrig, 2016)
SoMaJo current version
AIPHES (Remus et al., 2016)
COW (Schäfer/Bildhauer, 2012)
LTL-UDE (Horsmann/Zesch, 2016)
baseline

SoMaJo

Proisl and Uhrig (2016)

- Rule-based: cascade of regular expressions
- Can output the token class for each token, e.g. number, emoticon, abbreviation, etc.
- Can also output additional information for each token that can help to reconstruct the original untokenized text

```
echo 'der beste Betreuer? - >ProfSmith! : )' | somajo-tokenizer -cet -
der       regular
beste     regular
Betreuer  regular  SpaceAfter=No
?         symbol
->        symbol   SpaceAfter=No, OriginalSpelling="- >"
Prof      regular  SpaceAfter=No
Smith     regular  SpaceAfter=No
!         symbol
:)        emoticon OriginalSpelling=": )"
```

POS tagging – New tags added to STTS (1)

Beißwenger et al. (2015b)

- Tags for CMC phenomena:
  - EMO(ASC|IMG) (ASCII/graphic emoticon): `:-)` `^^` `O.O` ☺ ☹
  - AKW (interaction word): `*lach*, freu, grübel, *lol*`
  - HST (hash tag): `Kreta war super!  #Urlaub`
  - ADR (addressing term): `@lothar:  Wie isset so?`
  - URL: `https://www.uni-bamberg.de/germ-ling/`
  - EML (e-mail address): `thomas.proisl@fau.de`

- Tags for phenomena in colloquial registers:
  - Contractions with pronoun: (VV|VM|VA|KOUS|PPER)PPER: `schreibste, ichs, obse`
  - Contractions with article: APPRART, ADVART: `vorm, fürn, sone`

POS tagging – New tags added to STTS (1)

Beißwenger et al. (2015b)

- Tags for CMC phenomena:
  - EMO(ASC|IMG) (ASCII/graphic emoticon): `:-)` `^^` `O.O` ☺ ☹
  - AKW (interaction word): `*lach*, freu, grübel, *lol*`
  - HST (hash tag): `Kreta war super!  #Urlaub`
  - ADR (addressing term): `@lothar:  Wie isset so?`
  - URL: `https://www.uni-bamberg.de/germ-ling/`
  - EML (e-mail address): `thomas.proisl@fau.de`

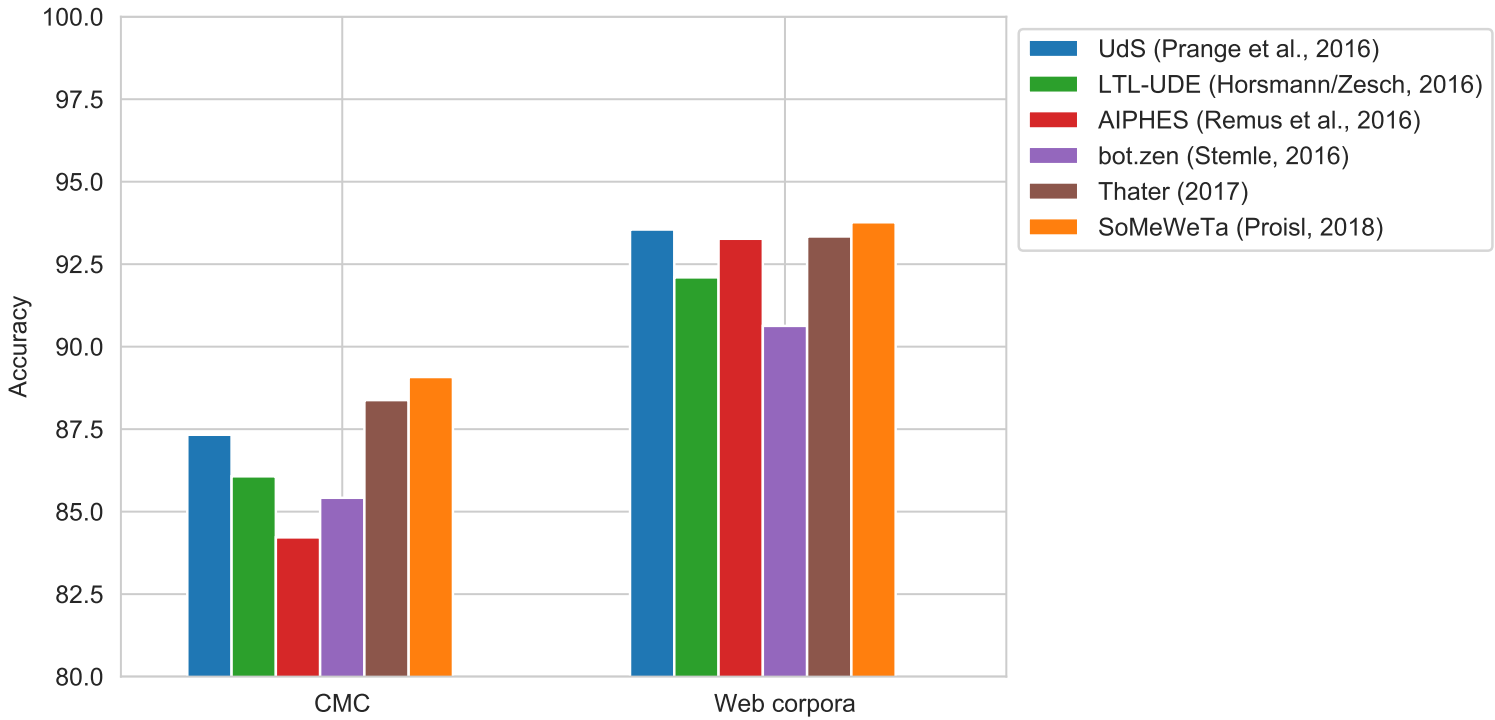
- Tags for phenomena in colloquial registers:
  - Contractions with pronoun: (VV|VM|VA|KOUS|PPER)PPER: `schreibste, ichs, obse`
  - Contractions with article: APPRART, ADVART: `vorm, fürn, sone`

POS tagging – New tags added to STTS (2)

Beißwenger et al. (2015b)

- Other refinements of STTS:
  - PTKIFG (intensifier/gradation particles): `sehr schön, voll geil`
  - PTKMA (modal particles and downtoners): `Das ist ja doof`
  - PTKMWL (particle as part of MW lexeme): `keine mehr, noch mal`
  - DM (discourse markers): `weil, obwohl, nur`
  - ONO (onomatopoeia): `boing, miau, zisch`
- PTK(IFG|MA|MWL) correspond to ADV or ADJD in standard STTS
- DM corresponds to KOUS or ADV

Results


100.0
97.5
95.0
92.5
90.0
87.5
85.0
82.5
80.0
Accuracy
CMC
Web corpora
UdS (Prange et al., 2016)
LTL-UDE (Horsmann/Zesch, 2016)
AIPHES (Remus et al., 2016)
bot.zen (Stemle, 2016)
Thater (2017)
SoMeWeTa (Proisl, 2018)

SoMeWeTa

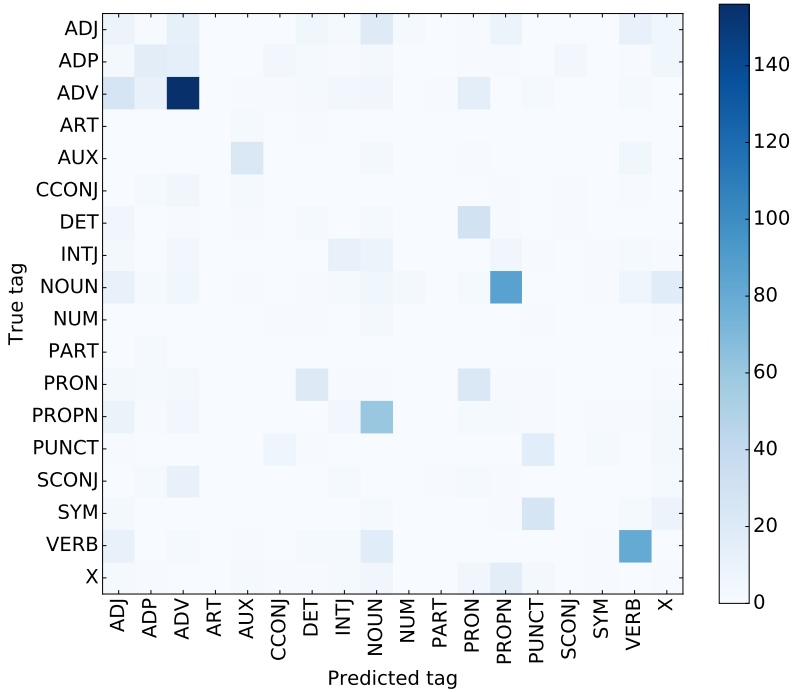
Proisl (2018)

- Averaged structured perceptron using beam search and early update strategy (Rosenblatt, 1958; Freund and Schapire, 1999; Collins, 2002; Collins and Roark, 2004)
- Domain adaptation (Chelba and Acero, 2004):
  - Use model trained on TIGER as prior on weights
- External resources:
  - Brown clusters (Brown et al., 1992) extracted from DECOW14 (Schäfer and Bildhauer, 2012; Schäfer, 2015)
  - Coarse-grained word class information from Morphy (Lezius, 2000)

Error analysis

SoMeWeTa (Proisl, 2018)

- Misclassifications related to new particle tags (PTKIFG, PTKMA, PTKMWL)
- Other major sources of errors:
  - confusion of NN and NE
  - misclassifications within verb tags



Lemmatization of German CMC data

Adding lemmatization to the EmpiriST corpus

- CMC-specific additions to TIGER morphology annotation scheme

- Two lemmatization strategies:
  1. Surface-oriented lemmatization (based on inflectional suffixes, retains non-standard orthographical features)
     - ★ `Grigfe` → `Grigf`
     - ★ `Soooo` → `soooo`
  2. Normalized lemmatization (correct obvious spelling errors, standard form of non-standard tokens)
     - ★ `Grigfe` → `Griff`
     - ★ `Soooo` → `so`

Guidelines

- For most of the new tags, the lemma is the surface form (EMOASC, EMOIMG, AKW, HST, ADR, URL, EML, PTKIFG, PTKMA, PTKMW, DM, ONO)

- Contractions: Lemma of contraction is lemma of first component (cf. APPRART)
  - `schreibste` → `schreiben`
  - `isses` → `sein`
  - `obse` → `ob`
  - `ichs` → `ich`
  - `hamwas` → `haben`

- The tricky part is the normalization!
  - `das/dass/daß`
  - `hinstelt`
  - `Disko(thek)` vs. `Disku(ssion)`
  - `bißt`
  - `jedenfall` (tagged as NN)

Guidelines

What's in a name? – heuristics for proper (nick-)names in chat protocols

- Established proper name: lemma is capitalised (`marc` for user marc30 → Marc)
- Non-established proper name: lemma = surface form (`sternchen` for user Nudelsuppenstern → sternchen)
- Problematic cases: correspondences to lexicalised words, short forms (`lanto` for user Lantonie), creative nicknames/ inflectional phenomena (username `stoeps` is "pluralised" to `stöpse` → `stöps`)

Not to be disregarded: motivational issues in the final phase (`ausgeliefert` → `ausgeliefern`)

Annotation

- Gold standard data independently lemmatized by four student annotators
- Unclear cases decided in group meetings with task organizers
- Agreement between annotators and gold standard (accuracy)
  - Drop from inter-annotator to annotator-gold due to post-annotation change in guidelines w.r.t. proper names

	Surface-oriented lemmata				Normalized lemmata			
	AJ	DW	EH	LR	AJ	DW	EH	LR
gold	93.64	92.87	93.73	93.67	93.10	92.82	93.80	93.46
AJ		96.08	96.54	96.50		96.00	96.28	95.92
DW			96.21	96.55			96.33	96.19
EH				96.89				96.70

Baseline systems

- Do-nothing baseline: Always return the word form
- Weak baseline: Given a word form and a POS tag, return the most frequent lemma from TIGER ∪ EmpiriST training data
  - Fallback 1: Ignore case
  - Fallback 2: Return word form
- Strong baseline: Apache OpenNLP maximum entropy lemmatizer

Baseline results (accuracy, ignoring case)

Baseline	surface-oriented	normalized
Do-nothing	71.63	70.73
Weak	83.90	83.19
Strong	87.50	85.97

→ Room for improvement!

1. Introduction

2. EmpiriST Corpus and Derived Tools
   - Tokenization
   - POS Tagging

3. Lemmatization
   - Guidelines
   - Annotation
   - Baseline systems

4. Ongoing project: Creation of a German Reddit corpus
   - Data Gathering
   - Corpus Creation
   - Annotation

5. Conclusion

And now for something completely different.

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: 33[4]

[1] https://www.alexa.com/topsites/countries/US, 16/9/2019
[2] https://www.similarweb.com/top-websites/united-states, 16/9/2019
[3] https://www.alexa.com/topsites/countries/DE, 16/9/2019
[4] https://www.similarweb.com/top-websites/germany, 16/9/2019

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

Corpus Cleanup

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

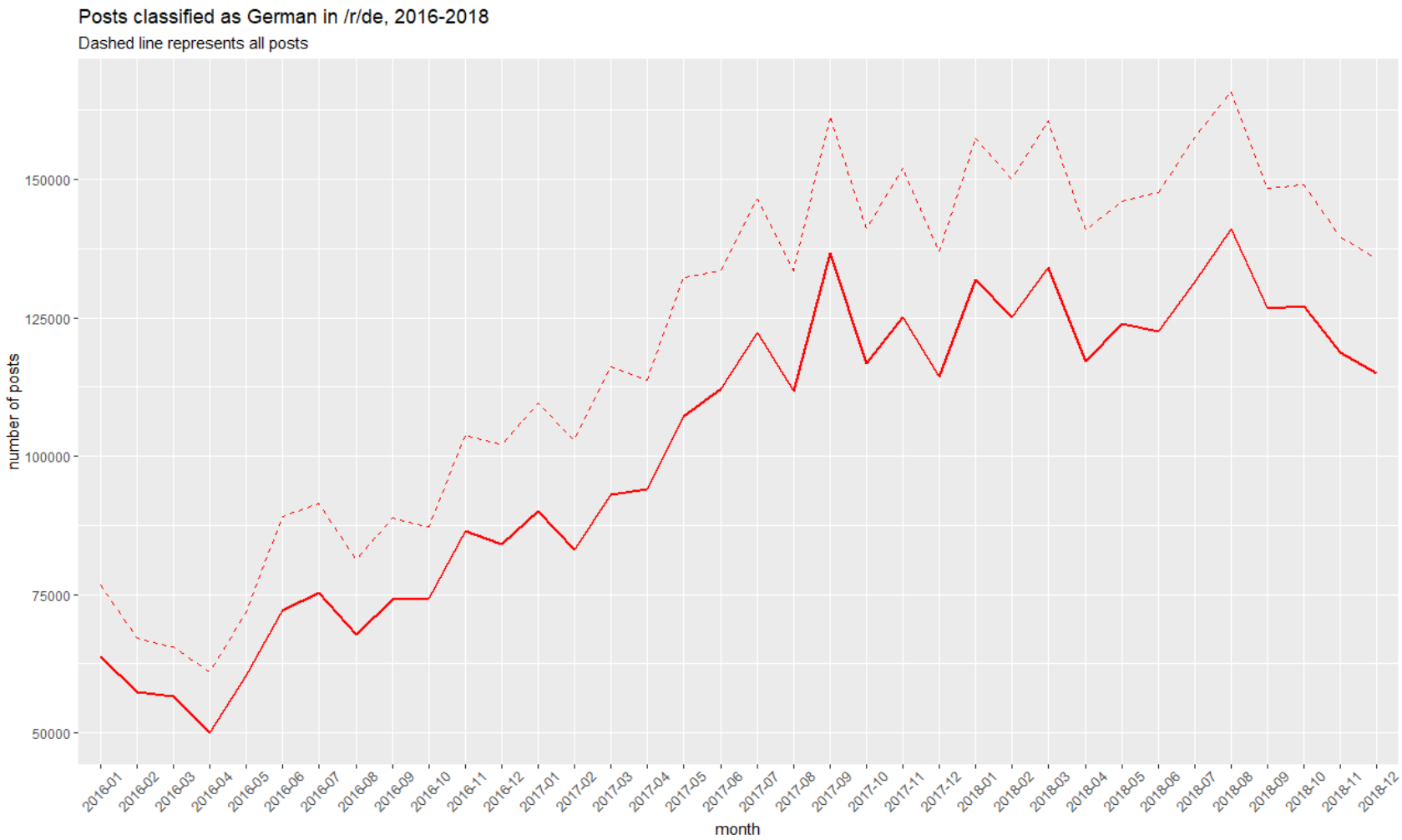
Subreddit	"German"	Total	Fraction
de	3644481	4364440	0.835
Austria	389041	483125	0.805
rocketbeans	142787	164699	0.867
AskReddit	125411	183147454	0.000685
edefreiheit	121370	147150	0.825
de_IAmA	58172	65768	0.885
Finanzen	48033	55009	0.873
Fireteams	38042	2125695	0.0179
soccer	33727	22918204	0.00147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

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

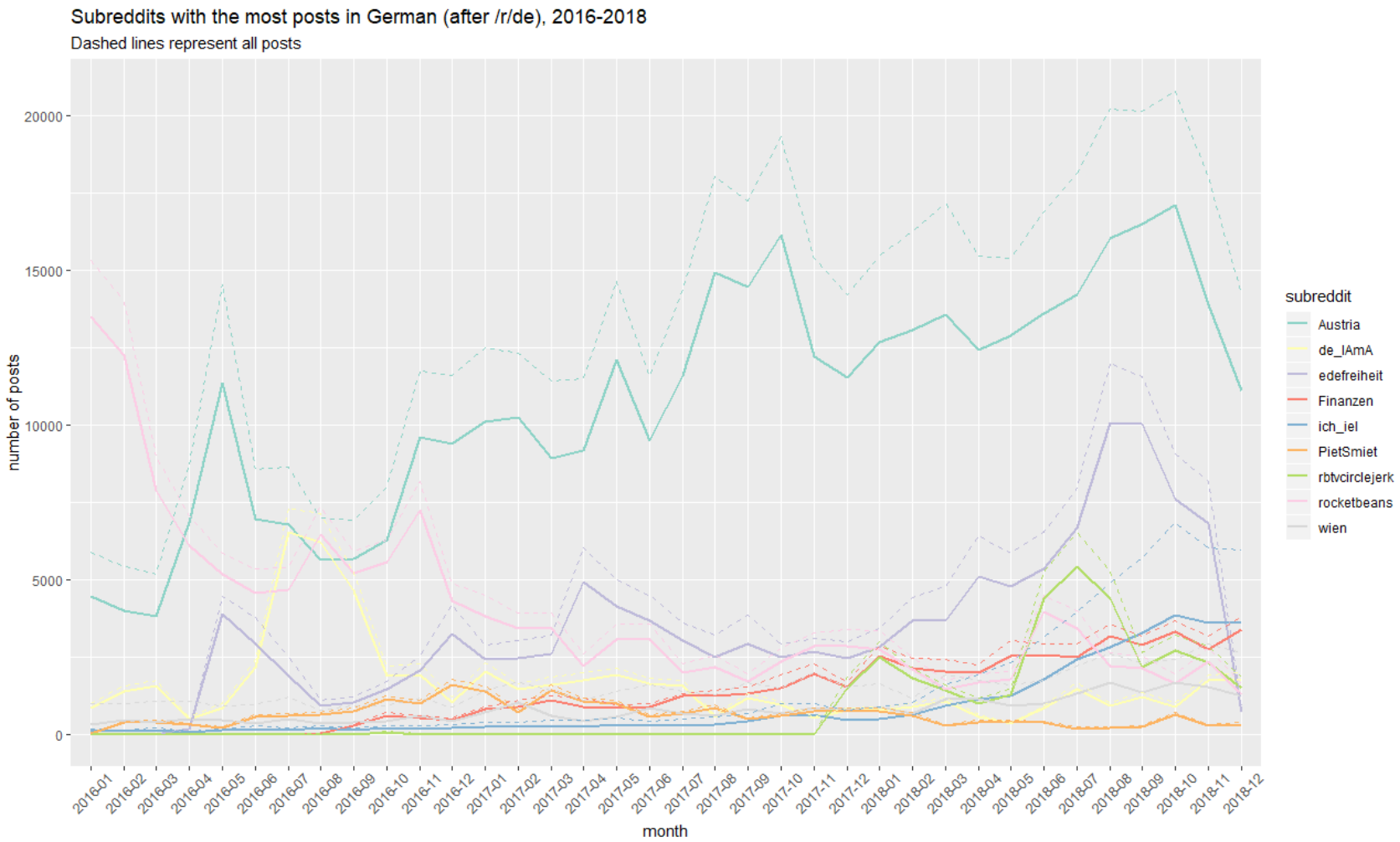
Corpus Cleanup

- 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: Emoticon as Punctuation

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
...				

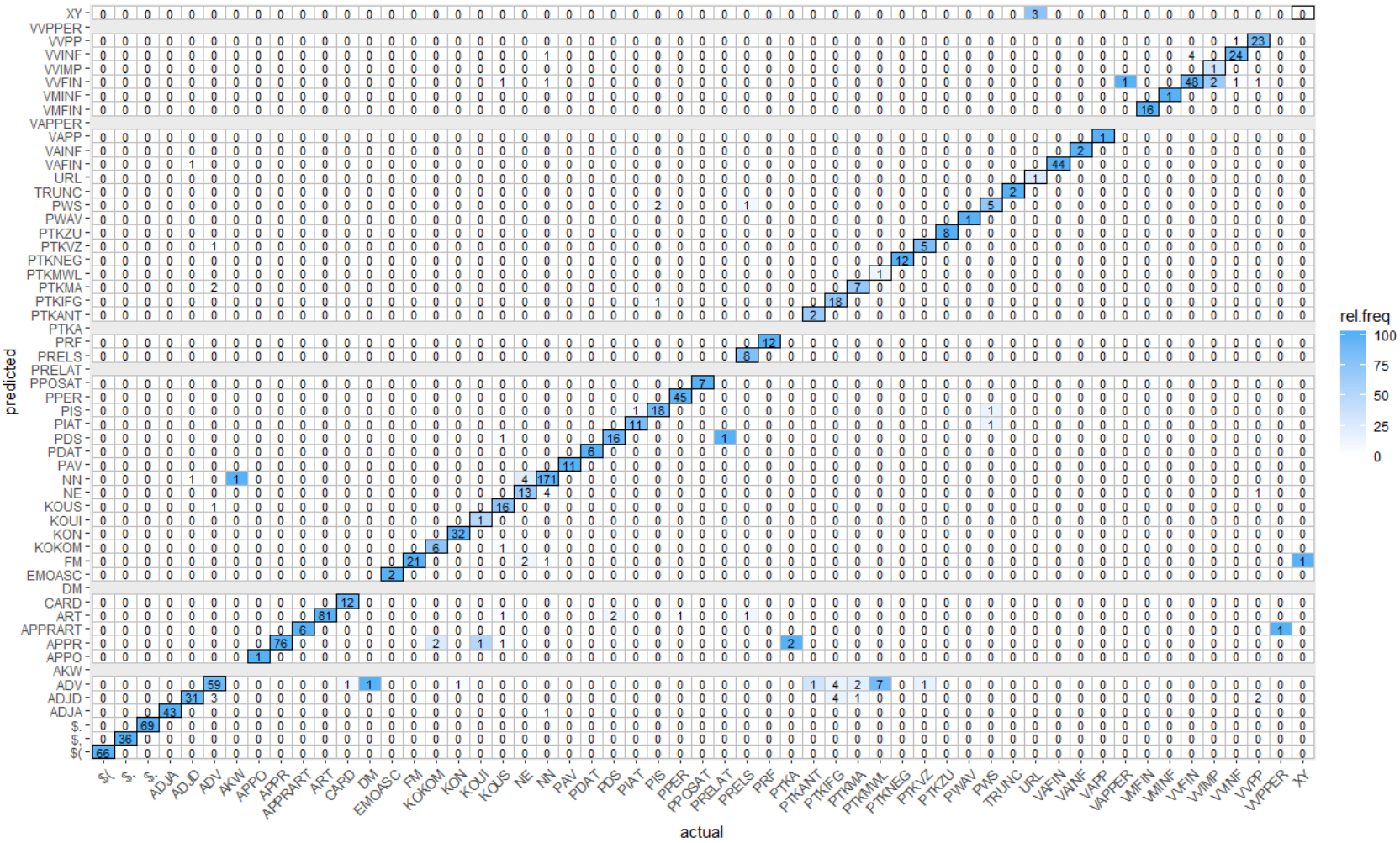
POS Tagging

- Random sample tagged with SoMeWeTa (trained on EmpiriST corpus)
- Corrected POS tags in a small number of comments (1186 tokens – so far) to evaluate performance
- Accuracy: 92.6%

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

Confusion Matrix



Conclusion

What we've learned

- Do not blindly use others' methods
- Pay close attention to tokenization and sentence splitting
- Adapt it if necessary
- STTS-IBK is a little weird (and we might need to discuss its rules and categories)

Reddit Corpus

- More corpus cleaning necessary
- Tokenization rules have to be updated
- POS-tagging and lemmatization further down the pipeline

Links

Tools

- SoMaJo:
  `https://github.com/tsproisl/SoMaJo`
- SoMeWeTa:
  `https://github.com/tsproisl/SoMeWeTa`

Data

- EmpiriST 2015:
  `https://sites.google.com/site/empirist2015/`
- EmpiriST corpus (with lemmata):
  `https://github.com/fau-klue/empirist-corpus`

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, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. Richtlinie für die manuelle Tokenisierung von Sprachdaten aus Genres internetbasierter Kommunikation. Guideline document, 2015a.

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, 2015b.

Michael Beißwenger, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. EmpiriST 2015: A shared task on the automatic linguistic annotation of computer-mediated communication and web corpora. 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 44–56, Berlin, 2016. Association for Computational Linguistics.

References II

Peter F. Brown, Vincent J. Della Pietra, Peter V. de Souza, Jennifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

Ciprian Chelba and Alex Acero. Adaptation of maximum entropy capitalizer: Little data can help a lot. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 285–292, Barcelona, 2004. Association for Computational Linguistics.

Michael Collins. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 1–8, Philadelphia, PA, 2002. Association for Computational Linguistics.

Michael Collins and Brian Roark. Incremental parsing with the perceptron algorithm. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pages 111–118, Barcelona, 2004. Association for Computational Linguistics.

Yoav Freund and Robert E. Schapire. Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277–296, 1999.

References III

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.

Tobias Horsmann and Torsten Zesch. LTL-UDE @ EmpiriST 2015: Tokenization and pos tagging of social media text. 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 120–126, Berlin, 2016. Association for Computational Linguistics.

Wolfgang Lezius. Morphy – German morphology, part-of-speech tagging and applications. In Ulrich Heid, Stefan Evert, Egbert Lehmann, and Christian Rohrer, editors, *Proceedings of the 9th EURALEX International Congress*, pages 619–623, Stuttgart, 2000. Institut für Maschinelle Sprachverarbeitung.

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 IV

Jakob Prange, Andrea Horbach, and Stefan Thater. UdS-(retrain|distributional|surface): Improving pos tagging for oov words in German cmc and web data. 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 63–71, Berlin, 2016. Association for Computational Linguistics.

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.

References V

Steffen Remus, Gerold Hintz, Chris Biemann, Christian M. Meyer, Darina Benikova, Judith Eckle-Kohler, Margot Mieskes, and Thomas Arnold. EmpiriST: AIPHES – Robust tokenization and pos-tagging for different genres. 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 106–114, Berlin, 2016. Association for Computational Linguistics.

Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

Roland Schäfer. Processing and querying large web corpora with the COW14 architecture. In Piotr Bański, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen, and Andreas Witt, editors, *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, pages 28–34, Lancaster, 2015. UCREL, IDS.

Roland Schäfer and Felix Bildhauer. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 486–493, Istanbul, 2012. European Language Resources Association.

References VI

Egon Stemle. bot.zen @ EmpiriST 2015 – A minimally-deep learning pos-tagger (trained for German cmc and web data). 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 115–119, Berlin, 2016. Association for Computational Linguistics.

Stefan Thater. Fine-grained POS tagging of German social media and web texts. In Georg Rehm and Thierry Declerck, editors, *Language Technologies for the Challenges of the Digital Age - 27th International Conference, GSCL 2017*, volume 10713 of *Lecture Notes in Computer Science*, pages 72–80. Springer, 2017.

Thanks for listening.
Questions?