EmpiriST Corpus 2.0: Adding Manual Normalization, Lemmatization and Semantic Tagging to a German Web and CMC Corpus

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

Abstract

The EmpiriST corpus (Beißwenger et al., 2016) is a manually tokenized and part-of-speech tagged corpus of approximately 23,000 tokens of German Web and CMC (computer-mediated communication) data. We extend the corpus with manually created annotation layers for word form normalization, lemmatization and lexical semantics. All annotations have been independently performed by multiple human annotators. We report inter-annotator agreements and results of baseline systems and state-of-the-art off-the-shelf tools.

Keywords: CMC, annotation, resources, normalization, lemmatization, semantic tagging

1. Introduction

Manually annotated data are crucial for training and evaluating statistical tools such as POS taggers and lemmatizers. The creation of these "gold standards" for corpus annotation layers is thus an inevitable (though labour-intensive and tedious) endeavour to make language data accessible. While there is a comparatively large amount of manually annotated data available for English, especially for standard (newspaper) texts, other languages and registers (such as computer-mediated communication, CMC) do not enjoy such great popularity. A notable exception are English Twitter data, for which both manually annotated corpora and designated tools have been developed (Ritter et al., 2011; Owoputi et al., 2013; Kong et al., 2014).

Our focus is on German web and CMC data. Off-the-shelf natural language processing (NLP) tools trained on newspaper corpora typically show a relatively poor performance on this kind of out-of-domain data (Giesbrecht and Evert, 2009; Neunerdt et al., 2013). As has been noted before, there are major linguistic differences between CMC and standard German (Haase et al., 1997; Runkehl et al., 1998; Dietterle et al., 2017; Beißwenger and Pappert, 2018): Computer-mediated communication, and chat communication in particular, has often been described as being "conceptually oral", i. e. exhibiting phenomena typically associated with oral communication. Examples include colloquial or dialectal word forms and constructions and utterances that are not syntactically well-formed. Another well-known phenomenon are substitutes for some of the non-verbal signals of oral communication, e. g. emoticons or action words (**freu**, from *(sich) freuen*, 'to rejoice'). Substitutes for stress and prosody include character repetitions, all caps or simple mark-up (surrounding a word or phrase with asterisks, slashes, underscores, etc.). A higher rate of spelling errors (sometimes due to production speed), deliberate creative spellings and the use of CMC-specific acronyms (*LOL*, *ROFL*, *IMHO*) are also often associated with CMC data.

In this paper, we describe our additions to the EmpiriST corpus (cf. Section 2.), a manually tokenized and part-of-speech tagged corpus of approximately 23,000 tokens of German Web and CMC data with subsequently added manually identified sentence boundaries. We added four manually created layers of annotation:

- normalized spelling
- surface-oriented lemma
- normalized lemma
- UCREL Semantic Analysis System (USAS) tag

Normalization of tokens and lemmata is a reasonable processing step for CMC data, since orthographic mistakes are ubiquitous. Lemmatization is crucial for general corpus indexing purposes as well as for many applications in lexicography, text classification, discourse analysis, etc. Just like lemmatization enables reasonable grouping of several words, semantic tags group together various related word senses, which can also be exploited e. g. for discourse analysis.

	CMC	Web	Total
Training	5,109	4,944	10,053
Test	5,237	7,568	12,805
Total	10,346	12,512	22,858

Table 1: Sizes of the EmpiriST corpus and its subsets in tokens.

2. The EmpiriST Corpus

The EmpiriST corpus is a manually annotated corpus consisting of German web pages and German computer-mediated communication (CMC), i. e. written discourse. Examples for CMC genres are monologic and dialogic tweets, social and professional chats, threads from Wikipedia talk pages, WhatsApp interactions and blog comments. Table 1 gives an overview of the sizes of the corpus and its subsets in tokens. The dataset was originally created by Beißwenger et al. (2016) as a gold standard for the EmpiriST 2015 shared task[1]

[1] https://sites.google.com/site/empirist2015/

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020) , pages 6144–6150
Marseille, 11–16 May 2020

and featured manual tokenization and part-of-speech tagging according to custom annotation guidelines. The tokenization guidelines (Beißwenger et al., 2015a)[2] cover a wide range of CMC-specific phenomena, including, for example, frequently used acronyms (*aka*, *cu*), typos and speed-writing phenomena (*schona ber* 'yesb ut', *maldrüber*), contracted forms (*machstes* from *machst es* or even *machst du es* 'make you it', *nochn* from *noch ein* 'another'), emoticons, hashtags, addressing terms, etc., and have been implemented, inter alia, by SoMaJo (Proisl and Uhrig, 2016)[3], the winning tokenizer of the shared task. For POS tagging, the STTS_IBK tag set (Beißwenger et al., 2015b)[2] has been used, which builds on the Stuttgart-Tübingen-Tagset (STTS; Schiller et al. (1999)) and extends it with tags for phenomena found in CMC genres (emoticons, hashtags, etc.) or in spontaneous spoken or conceptually oral language (e. g. various types of contractions). Pretrained tagger models for STTS_IBK are available, inter alia, for SoMeWeTa (Proisl, 2018)[4], GermaPOS (Remus et al., 2016)[5] and the LTL-UDE system (Horsmann and Zesch, 2016)[6].

Subsequently, Rehbein et al. (2018) manually added sentence boundaries to the EmpiriST corpus, automatically mapped the part-of-speech tags to UD POS tags (Nivre et al., 2016)[7] and incorporated the dataset into their harmonised testsuite for POS tagging of German social media data.[8] For the identification of sentence boundaries, they used the following rules to guide the segmentation:

> - Hashtags and URLs at the beginning or the end of the tweet that are not integrated in the sentence are separated and form their own unit [. . .].
> - Emoticons are treated as non-verbal comments to the text and are thus integrated in the utterance.
> - Interjections (*Aaahh*), inflectives (**grins**), fillers (*ähm*) and acronyms typical for CMC (*lol*, *OMG*) are also not separated but considered as part of the message.
>
> (Rehbein et al., 2018, p. 20)

The current version of the corpus includes both the sentence boundaries and the UD POS tags.

3. New Annotation Layers in Version 2.0

For version 2.0, we converted the EmpiriST corpus into a corpus linguistic standard format and manually created annotation layers for word form normalization, lemmatization and lexical semantics.

The annotated corpus is freely available under a Creative Commons license and can be found under `https://github.com/fau-klue/empirist-corpus` along with information on our lemmatization guidelines.

[2] `https://sites.google.com/site/empirist2015/home/annotation-guidelines`

[3] `https://github.com/tsproisl/SoMaJo`

[4] `https://github.com/tsproisl/SoMeWeTa`

[5] `https://github.com/AIPHES/GermaPOS`

[6] `https://github.com/Horsmann/EmpiriSharedTask2015`

[7] `https://universaldependencies.org/u/pos/all.html`

[8] `https://www.cl.uni-heidelberg.de/~rehbein/tweeDe.mhtml`

	AJ	DW	EH	LR
gold	94.45	93.85	94.42	94.23
AJ		98.11	98.09	98.04
DW			98.24	98.15
EH				98.20

Table 2: Agreement scores for normalization (case sensitive, accuracy).

3.1. Format Changes

Originally, the corpus was organized as a collection of text files with standalone tags marking the beginning of a new text or posting. We converted it into the "vertical" format used by the Open Corpus Workbench, CQPweb, SketchEngine, and similar corpus tools. i. e. a CoNLL-style format with tab-separated columns for token-level annotation and structural XML tags for texts, postings and sentences (cf. the example in Figure 1).

3.2. Normalization

CMC data often deviate from the norms of the written standard language and are conceptually closer to spoken language. This affects syntax and lexical choices but also spelling. Phenomena leading to non-standard spellings include contractions (*gehts* (= *geht es* 'goes it'), *sone* (= *so eine* 'such a')), elisions (*ne* (= *eine* 'a'), *hinziehn* (= *hinziehen* 'drag on')), creative spellings (*ver3fachte* (= *verdreifachte* 'tripled')), emphasis via character repetitions (*dahaaaa* (= *da* 'there'), *geeeil* (= *geil* 'cool, wicked')) and of course typos.

In our normalization efforts, we correct obvious typos (*das/dass*, *hinstelt* → *hinstellt* 'places, puts', *Grigfe* → *Griffe* 'grips, handles'), normalize to "new" (i. e. post spelling reform) spellings (*muß* → *muss* 'must') and generally normalize non-lexicalized forms to established standard forms (*hund* → *Hund* 'dog', *zB* → *z.B.* 'e. g.', *uuuh* → *uh*, *nen* → *einen*, *Disku* → *Diskussion* 'discussion'). For the complete guidelines (in German), see Proisl et al. (2019)[9].

The whole corpus was independently normalized by four student helpers. Unclear cases were decided in group meetings with the authors. Table 2 shows the agreement scores between the annotators and the adjudicated gold standard.[10]

In a relatively late stage of the annotation process, we changed the normalization and lemmatization guidelines for proper names. The subsequent changes to the adjudicated gold standard explain the lower agreement scores between the individual annotators and the final gold standard. Without these changes, the mean inter-annotator agreement score is 98.14; agreement with the prior version of the gold standard would obviously also be higher.

3.3. Lemmatization

In order to accommodate different users' needs, we implemented two different lemmatization strategies: Surface-oriented lemmatization and normalized lemmatization.

[9] `https://github.com/fau-klue/empirist-corpus/blob/master/doc/Lemmatisierungsrichtlinien.pdf`

[10] Values of Cohen's κ are practically the same.

```
<posting id="cmc_train_003_099" author="quaki" origid="1-114">
<s>
die         ART     DET     Z5        die         der       der
viecha      NN      NOUN    L2        Viecher     Viech     Viech
reissen     VVFIN   VERB    A1.1.2    reißen      reissen   reißen
imma        ADV     ADV     T1.1      immer       imma      immer
die         ART     DET     Z5        die         der       der
müllsäcke   NN      NOUN    O2        Müllsäcke   Müllsack  Müllsack
auf         PTKVZ   PART    A10       auf         auf       auf
hmmmm       ITJ     INTJ    Z4        hm          hmmmm     hm
</s>
</posting>
```

Figure 1: A one-sentence posting ('The critters always rip open the garbage bags, hm') illustrating the corpus format. The seven columns are: Word form, STTS_IBK tag, UD POS tag, USAS tag, normalized form, surface-oriented lemma, normalized lemma.

Surface-oriented lemmata are mainly based on the inflectional suffixes of the token and as far as possible, retain any non-standard orthographic features of the token. Possible use cases for these lemmata include the evaluation of affix-based lemmatization tools or studies on linguistic variation (e. g. by retaining the difference between colloquial and standard variants of high-frequency items). For normalized lemmata, on the other hand, obvious spelling errors are corrected and non-standard forms are treated as standard forms. Normalized lemmatization is based on the normalized word forms (cf. previous section) and creates, as far as possible, standard German lemmata.

Surface-oriented lemmatization treats deviations from the standard as creative language use. For example, the mis-spelled word form *Grigfe*, tagged as NN (noun), is treated as the plural of a non-lexicalized noun *Grigf* (whereas normalized lemmatization is based on the normalized word form *Griffe* 'grips, handles' and results in the lemma *Griff*). Similarly, the mis-spelled word form *hinstelt*, tagged as VVFIN (finite full verb), is treated as an inflected form of a newly created prefix verb *hinstelen*, which might be derived from the noun *Stele* 'stele' (whereas normalized lemmatization based on the corrected word form *hinstellt* results in the lemma *hinstellen* 'place, put'). If inflectional suffixes are not sufficient, e. g. due to stem changes, surface-oriented lemmatization falls back to normalized lemmatization. Therefore, word forms like *iest* or *fannd* receive standard language lemmata, i. e. *sein* 'be' or *finden* 'find'.

Lemmatization follows the TIGER lemmatization guidelines (Crysmann et al., 2005)[11], to which we make extensions that cover the new POS tags introduced in STTS_IBK (Proisl et al., 2019)[9]. For most of the new tags, the lemmatization rules should not be too controversial because they cover tokens that do not inflect anyway (e. g. emoticons, email addresses, URLs, particles). For the new POS tags covering contracted forms, we proceed in analogy to the APPRART tag (contraction of preposition and article) and choose as lemma of the whole contraction the lemma of its first constituent.

The whole corpus was independently lemmatized according to both strategies by four student helpers. Unclear cases were decided in group meetings with the authors. Table 3 shows the agreement scores between the annotators and the adjudicated gold standard for surface-oriented lemmata and Table 4 for normalized lemmata.

	AJ	DW	EH	LR
gold	93.64	92.87	93.73	93.67
AJ		96.08	96.54	96.50
DW			96.21	96.55
EH				96.89

Table 3: Agreement scores for surface-oriented lemmatization (case sensitive, accuracy).

	AJ	DW	EH	LR
gold	93.10	92.82	93.80	93.46
AJ		96.00	96.28	95.92
DW			96.33	96.19
EH				96.70

Table 4: Agreement scores for normalized lemmatization (case sensitive, accuracy).

As explained in the previous section, the lower agreement scores between the individual annotators and the final gold standard are due to late-stage changes in the annotation guidelines with respect to proper names. Without these changes, the mean inter-annotator agreement scores are 96.46 for surface-oriented lemmatization and 96.24 for normalized lemmatization.

3.4. Semantic Tagging

Token-level semantic tags were added using the USAS tagset.[12] USAS features 21 broad domains intended to capture the most important aspects of lexical semantics in everyday language (*A: Abstract and General Terms*, *Y: Science and Technology* etc.). Most of these categories can be sub-divided into several levels of abstraction (*A1: General Terms*; *A1.1.1: General Actions, making etc.*; *A1.1.2: Damaging and destroying*). The tagset has been applied to

[11] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/TIGERCorpus/annotation/tiger_scheme-morph.pdf

[12] The English version of the full tagset (Archer et al., 2002) can be found at http://ucrel.lancs.ac.uk/usas/usas%20guide.pdf

	coarse tags	fine tags
total agreement	86.5%	78%
partial agreement	4.7%	3.1%
different tags	13.5%	22%

Table 5: Agreement scores for semantic tagging.

multiple languages, with the present contribution being its first application to German data. The data were annotated independently by one of the authors and a student assistant. Semantic tagging has yielded new possibilities for applications like the semi-automatic identification of metaphor in corpora (Demmen et al., 2015; Potts and Semino, 2017).
At the time of writing, the development of guidelines for semantic tagging is an ongoing effort. Preliminary agreement scores between the annotators developing the gold standard are provided in Table 5 for the first 7,000 tokens of the corpus; amounting to roughly a third of the entire dataset. The coarse tags refer to the general category (A–Z), while the fine tags correspond to the finer sub-set of that category – i. e. the granularity that was actually used for annotation. Partial agreement was defined as one tag being the same, as each token may have two (and occasionally more) tags:
Initial agreement scores leave room for improvement; as is to be expected for such a complex task. However, the coarse categories score markedly better than the fine-grained ones. While it is obvious that the score for the smaller tag-set will be higher, some coarse domains are highly abstract, with some conceptual overlap, and it is encouraging that in the majority of cases, the general assessment of semantic category has been similar. In many cases, differences between the annotators are systematic and related to differing assessments of the category scope. For instance, prepositions were tagged as grammatical markers by one annotator and as *Location and direction* by the other.
Moreover, several areas of CMC phenomena resulted in a need for developing workarounds in our gold standard, as the tag-set is not inherently designed to reflect them. While the adjudication is still ongoing, our current guidelines propose the following solutions:

- Action words like **freu** (*happy*) are treated as the combination of an emotion or action tag and a second tag denoting a speech act
- The same holds true for emoticons: While :) does not readily fit the existing category scheme, it is currently treated as *E4.2 Happy/sad: Contentment* and *Q2.2 Speech Acts*
- Media that had not been invented when the tagset was developed require similar heuristics: we therefore treat *Blog* as *Q1.2 Paper documents and writing* + *Y2 Information technology and computing*

4. Baselines and Experiments

Average human performance is around 98.1% accuracy on the normalization task and between 96.2% and 96.5% on the two lemmatization tasks. To get a more complete picture of these tasks, we implemented two baseline systems and evaluated two state-of-the-art lemmatizers for German in an off-the-shelf setting. The results are summarized in Table 6. In addition, we experimented with a finite-state morphological analyzer and two statistical lemmatizers. The two baseline systems and our wrapper script for the morphological analyzer can be found in the repository.[13]

4.1. Baseline Systems

We evaluated two different baseline strategies, using the test subset of the annotated corpus (see Table 1) as evaluation data set: A do-nothing strategy and a simple lookup-based strategy. The do-nothing normalizer and lemmatizer simply return the original word form. Since most word forms in the corpus use standard orthography, this strategy results in 91.28% accuracy on the normalization task. Not surprisingly, the strategy works less well for lemmatization. However, almost two thirds of the tokens are not inflected and therefore get correctly "lemmatized" by this baseline strategy.
For the lookup-based strategy, we take the word form (in its original casing) and the gold POS tag and return the most frequent lemma or normalized word form that we observe for that combination in a manually annotated reference corpus (cf. Table 6). If there are no matches for a given word-POS combination, we repeat the process ignoring case. The final fallback is to return the original word form.
The lookup-based strategy is remarkably effective: By looking up normalized word forms in the training set of the EmpiriST corpus and lemmata in the union of the TIGER corpus and the EmpiriST training set, the baseline system achieves 96.09% accuracy on the normalization task and accuracies of 94.52% and 93.92% on the two lemmatization tasks.

4.2. Off-the-Shelf Tools

According to a recent evaluation (Ortmann et al., 2019), the two best-performing tools for lemmatizing German text are RNNTagger (Schmid, 2019) and TreeTagger (Schmid, 1994; Schmid, 1995). Both tools do their own part-of-speech tagging and we evaluate them using their own predicted tags instead of the gold tags.
One problem with evaluating lemmatizers is that they can adhere to different lemmatization guidelines. While the lemmatization component of RNNTagger is trained on the TIGER corpus and produces lemmata that are compatible to our gold standard, TreeTagger follows slightly different conventions which leads to a weak performance out of the box (accuracies of 80.80% and 80.34%). A brief analysis suggests that the major differences are the treatment of articles (TIGER lemmatizes them to *ein* and *der*, TreeTagger to *eine* and *die*), contractions (TIGER and our guidelines use the first component, e. g. *im* (= *in dem* 'in the') → *in*, TreeTagger produces a complex lemma, e. g. *im* → *in+die*), as well as cardinal and ordinal numbers (TIGER uses the surface form, TreeTagger assigns the pseudo-lemmata @*card*@ and @*ord*@). Fixing these differences in a search-and-replace postprocessing step drastically increased the accuracies to 92.01% and 91.74%, almost to the level of RNNTagger (92.71% and 92.06%).

[13] https://github.com/fau-klue/empirist-corpus/tree/master/baselines

Strategy	Normalization	Lemmatization		Unknown words
		Surface-oriented	Normalized	
Use word form	91.28	66.18	65.44	–
Lookup EmpiriST	**96.09**	85.75	85.22	34.26%
Lookup TIGER	91.28	93.27	92.53	23.77%
Lookup EmpiriST + TIGER	**96.09**	**94.52**	**93.92**	13.78%
TreeTagger	–	80.80	80.34	9.82%
TreeTagger + postproc.	–	92.01	91.74	9.82%
RNNTagger	–	92.71	92.06	–
SMOR	–	74.01	74.04	11.12%
SMOR + postproc.	–	89.21	89.22	9.30%
SMOR + postproc. + heuristics	–	**96.96**	**96.20**	–

Table 6: Performance of baseline systems, off-the-shelf tools, and the SMOR wrapper script (case sensitive, accuracy). Where applicable, we indicate the proportion of unknown words.

One notable difference between TreeTagger and RNNTagger is how they treat unknown words. While TreeTagger simply uses the word form, RNNTagger tries to lemmatize all words, including unknown words, even if it does not make sense, e. g. for URLs.[14]

At first sight, it might be surprising that neither of the two taggers is able to beat the lookup-based baseline strategy. However, we need to keep in mind that the two tools have not been exposed to CMC phenomena during training, i. e. they are of course not magically able to lemmatize these phenomena according to our guidelines. Another important difference is that the baselines make use of the gold tags whereas TreeTagger and RNNTagger base their lemmata on their own predicted tags which are only 87.04% and 86.61% correct.[15] Finally, RNNTagger suffers somewhat from attempting to lemmatize non-inflected tokens.

4.3. Morphological Analysis

SMOR (Schmid, 2004) is a finite-state morphological analyzer for German, which is freely available for non-commercial purposes.[16] SMOR provides a lemmatization component, which maps word forms to combinations of STTS part-of-speech tag and corresponding lemma. Unlike the tools evaluated in Section 4.2., the SMOR lemmatizer cannot be used off-the-shelf because it does not recognize capitalized words at the beginning of a sentence and has incomplete coverage of punctuation and other non-words. We implemented a small Perl script which automatically looks up different capitalizations if a word form is not recognized immediately and keeps punctuation and other non-words unchanged as lemma (including `ADR`, `HST` and `URL`). The script also corrects some minor differences between STTS and the POS tags generated by the SMOR lemmatizer. In all our experiments, lemmatization is based on the gold standard POS tags: an SMOR analysis will always be ignored if its POS tag doesn't match the tag in the corpus.

With this minimal wrapper, SMOR only achieves an accuracy of 74.01% for surface-oriented lemmatization (see Table 6 for results on normalized words). This is partly due to a fairly high proportion of 11.12% unknown words, which are considered as lemmatization errors. A much bigger factor are systematic differences in lemmatization conventions between SMOR and TIGER, which affect most closed-class words. With a post-processing step that uses mappings for closed-class words obtained from the TIGER corpus,[17] accuracy increases to 89.21%.

The proportion of unknown words is still relatively high (9.30%), but SMOR is highly reliable on known words with an accuracy of 98.36%. Finally, we added the standard heuristic of inserting the surface form as lemma for unknown words (with some case normalization). This version of SMOR outperforms all other approaches with a lemmatization accuracy of 96.96%. The remaining lemmatization errors show no obvious systematic patterns.

4.4. Statistical Lemmatizers

We experimented with two statistical lemmatizers: Apache OpenNLP[18] and mate-tools (Björkelund et al., 2010)[19]. We trained the lemmatizers once on the training subset of the EmpiriST corpus and once on the union of the EmpiriST training set and the TIGER corpus and evaluated them

[14] For example, the proposed lemma for the URL `http://www.youtube.com/watch?v=2w1g-idt-8U` is `http://www.youtube.com/wat-idt-80`.

[15] Evaluated using the mapping from STTS_IBK to STTS 1.0 specified by Beißwenger et al. (2016, p. 53).

[16] `https://www.cis.uni-muenchen.de/~schmid/tools/SMOR/`

[17] Mappings take the form of lookup tables for articles, adpositions, conjunctions and pronouns, obtained directly from the TIGER corpus. The first lookup table has 933 entries of the form (lowercased word form, POS tag) $\mapsto$ TIGER lemma, covering 677 word forms. A second lookup table attempts to adjust the raw SMOR lemmatization, with 121 entries of the form (SMOR lemma, POS tag) $\mapsto$ TIGER lemma, covering 91 SMOR lemmata. For both tables, filtering heuristics had to be applied because of lemmatization ambiguities not resolved by the POS tags and because of inconsistencies in the TIGER annotation. Note that in contrast to the best baseline system described in Section 4.1., the EmpiriST training corpus was not used at all.

[18] `https://opennlp.apache.org/`

[19] `https://code.google.com/archive/p/mate-tools/`

Tool	Case-sensitive		Case-insensitive	
	Surface-oriented	Normalized	Surface-oriented	Normalized
OpenNLP EmpiriST	76.17	75.81	92.78	92.01
OpenNLP TIGER + EmpiriST	**78.51**	**78.13**	**97.51**	**96.84**
Mate-tools EmpiriST	71.00	70.55	86.36	85.67
mate-tools TIGER + EmpiriST	76.83	76.26	94.30	93.50

Table 7: Evaluation results for the statistical lemmatizers.

against the test subset of the EmpiriST corpus. The results of the case-sensitive evaluation are rather disappointing (Table 7). Both OpenNLP and mate-tools perform worse than our lookup-based baseline system. A closer look at the output of the two systems showed that both did not learn to output capitalized lemmata. Therefore, we also performed a case-insensitive evaluation. The results show that OpenNLP could even outperform SMOR if it were combined with a suitable post-processing step to correct the capitalization of its output.

5. Conclusion and Future Work

We presented an updated version of the EmpiriST corpus with new annotation layers containing normalized word forms, two different kinds of lemmata and semantic tags. Human performance is 98.1% accuracy on the normalization task and between 96.2% and 96.5% on the two lemmatization tasks. The simple baselines we implemented are within two percentage points of human performance, a more sophisticated approach based on a finite-state morphological analyzer even surpasses human performance.

In the future, we would like to extend the corpus with additional data, e. g. from Reddit and Twitter, and to add further annotation layers, e. g. for named-entity recognition, semantic role labeling or syntactic analysis. We also plan to provide alternative lemmatizations for prefix verbs and contractions. In the case of prefix verbs, the usual practice is somewhat inconsistent. When a verb (e. g. *nachmachen* 'imitate, reproduce') is not split (e. g. *nachgemacht*), the lemma includes the prefix (*nachmachen*). However, when it is split in a sentence (e. g. *macht ... nach*), two different lemmata are assigned (*machen* and *nach*), making it more difficult to retrieve all instances of the prefix verb in a corpus. Ideally, the same lemma would be assigned to the verb in both cases. In the case of contractions, lemmatization currently results in a loss of information, since only the lemma of the contraction's first component is retained (cf. Section 3.3.). An alternative way of doing this (without changing the tokenization) would be to combine the lemmata of all components. Thus, the lemma of *machstes* would not only be *machen* but *machen+du+es*.

6. Bibliographical References

Archer, D., Wilson, A., and Rayson, P., (2002). *Introduction to the USAS category system*. UCREL, Lancaster University.

Beißwenger, M. and Pappert, S. (2018). Internetbasierte Kommunikation. In Frank Liedtke et al., editors, *Handbuch Pragmatik*, pages 448–459. J.B. Metzler, Stuttgart.

Beißwenger, M., Bartsch, S., Evert, S., and Würzner, K.-M. (2015a). Richtlinie für die manuelle Tokenisierung von Sprachdaten aus Genres internetbasierter Kommunikation. Guideline document.

Beißwenger, M., Bartz, T., Storrer, A., and Westpfahl, S. (2015b). 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.

Björkelund, A., Bohnet, B., Hafdell, L., and Nugues, P. (2010). A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Demonstrations Volume*, pages 33–36.

Crysmann, B., Hansen-Schirra, S., Smith, G., and Ziegler-Eisele, D. (2005). TIGER Morphologie-Annotationsschema. Guideline document.

Demmen, J., Semino, E., Demjen, Z., Koller, V., Hardie, A., Rayson, P., and Payne, S. (2015). A computer-assisted study of the use of violence metaphors for cancer and end of life by patients, family carers and health professionals. *International Journal of Corpus Linguistics*, 20(2):205–231.

Dietterle, B., Lüdeling, A., and Reznicek, M. (2017). Zur Syntax in Plauderchats. In Michael Beißwenger, editor, *Empirische Erforschung internetbasierter Kommunikation*, pages 47–80. De Gruyter, Berlin.

Giesbrecht, E. and Evert, S. (2009). Is part-of-speech tagging a solved task? An evaluation of POS taggers for the web as corpus. In Inaki Alegria, et al., editors, *Proceedings of the Fifth Web as Corpus Workshop (WAC5)*, pages 27–35, San Sebastian.

Haase, M., Huber, M., Krumeich, A., and Rehm, G. (1997). Internetkommunikation und Sprachwandel. In Rüdiger Weingarten, editor, *Sprachwandel durch Computer*, pages 51–85. VS Verlag für Sozialwissenschaften, Wiesbaden.

Horsmann, T. and Zesch, T. (2016). LTL-UDE @ EmpiriST 2015: Tokenization and pos tagging of social media text. In Paul Cook, et al., editors, *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 120–126, Berlin. Association for Computational Linguistics.

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., and Smith, N. A. (2014). A dependency parser for tweets. In *EMNLP 2014*, pages 1001–1012, October.

Neunerdt, M., Trevisan, B., Reyer, M., and Mathar, R. (2013). Part-of-speech tagging for social media texts. In Iryna Gurevych, et al., editors, *Language Processing and Knowledge in the Web*, pages 139–150, Berlin, Heidelberg. Springer Berlin Heidelberg.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož. European Language Resources Association.

Ortmann, K., Roussel, A., and Dipper, S. (2019). Evaluating off-the-shelf NLP tools for German. In *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 212–222, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *NAACL-HLT 2013*, pages 380–390.

Potts, A. and Semino, E. (2017). Healthcare professionals' online use of violence metaphors for care at the end of life in the us: a corpus-based comparison with the UK. *Corpora*, 12(1):55–84.

Proisl, T. and Uhrig, P. (2016). SoMaJo: State-of-the-art tokenization for German web and social media texts. In Paul Cook, et al., editors, *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin. Association for Computational Linguistics.

Proisl, T., Dykes, N., Heinrich, P., Kabashi, B., and Evert, S. (2019). Lemmatisierungsrichtlinien. Guideline document.

Proisl, T. (2018). SoMeWeTa: A part-of-speech tagger for German social media and web texts. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 665–670, Miyazaki. European Language Resources Association.

Rehbein, I., Ruppenhofer, J., and Zimmermann, V. (2018). A harmonised testsuite for POS tagging of German social media data. In Adrien Barbaresi, et al., editors, *KONVENS 2018 – Proceedings of the 14th Conference on Natural Language Processing, Vienna, Austria, September 19-21, 2018*, pages 18–28. Österreichische Akademie der Wissenschaften.

Remus, S., Hintz, G., Biemann, C., Meyer, C. M., Benikova, D., Eckle-Kohler, J., Mieskes, M., and Arnold, T. (2016). EmpiriST: AIPHES – Robust tokenization and postagging for different genres. In Paul Cook, et al., editors, *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 106–114, Berlin. Association for Computational Linguistics.

Ritter, A., Clark, S., Mausam, and Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA. Association for Computational Linguistics.

Runkehl, J., Schlobinski, P., and Siever, T. (1998). *Sprache und Kommunikation im Internet*. Westdeutscher Verlag, Opladen.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Technical report, IMS Stuttgart, SfS Tübingen.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the EACL SIGDAT-Workshop*, pages 47–50, Dublin.

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

7. Language Resource References

Apache Software Foundation. (2004–2020). *Apache OpenNLP*. `https://opennlp.apache.org/`.

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/`.

Bernd Bohnet and Anders Björkelund. (2014). *mate-tools*. `https://code.google.com/archive/p/mate-tools/`.

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, `https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger/`.

Ines Rehbein and Josef Ruppenhofer and Victor Zimmermann. (2018). *A harmonised testsuite for POS tagging of German social media data*. `https://www.cl.uni-heidelberg.de/~rehbein/tweeDe.mhtml`.

Helmut Schmid. (1994). *TreeTagger*. `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`.

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

Helmut Schmid. (2019). *RNNTagger*. `http://www.cis.uni-muenchen.de/~schmid/tools/RNNTagger/`.