EmpiriST 2015: A Shared Task on the Automatic Linguistic Annotation of Computer-Mediated Communication and Web Corpora

Michael Beißwenger
Universität Duisburg-Essen
45127 Essen, Germany
michael.beisswenger
@uni-due.de

Sabine Bartsch
Technische Universität Darmstadt
64293 Darmstadt
bartsch
@linglit.tu-darmstadt.de

Stefan Evert
FAU Erlangen-Nürnberg
91054 Erlangen, Germany
stefan.evert@fau.de

Kay-Michael Würzner
Berlin-Brandenburgische
Akademie der Wissenschaften
10405 Berlin, Germany
wuerzner@bbaw.de

Abstract

This paper describes the goals, design and results of a shared task on the automatic linguistic annotation of German language data from genres of computer-mediated communication (CMC), social media interactions and Web corpora. The two subtasks of tokenization and part-of-speech tagging were performed on two data sets: (i) a genuine *CMC data set* with samples from several CMC genres, and (ii) a *Web corpora data set* of CC-licensed Web pages which represents the type of data found in large corpora crawled from the Web. The teams participating in the shared task achieved a substantial improvement over current off-the-shelf tools for German. The best tokenizer reached an F_1-score of 99.57% (vs. 98.95% off-the-shelf baseline), while the best tagger reached an accuracy of 90.44% (vs. 84.86% baseline). The gold standard (more than 20,000 tokens of training and test data) is freely available online together with detailed annotation guidelines.

1 Motivation, premises and goals

Over the past decade, there has been a growing interest in collecting, processing and analyzing data from genres of computer-mediated communication and social media interactions (henceforth referred to as CMC) such as chats, blogs, forums, tweets, newsgroups, messaging applications (SMS, WhatsApp), interactions on "social network" sites and on wiki talk pages. The development of resources, tools and best practices for automatic linguistic processing and annotation of CMC discourse has turned out to be a desideratum for several fields of research in the humanities:

1. Large corpora crawled from the Web often contain substantial amounts of CMC (blogs, forums, etc.) and similar forms of non-canonical language. Such data are often regarded as "bycatch" that proves difficult for linguistic annotation by means of standard natural language processing (NLP) tools that are optimized for edited text (Giesbrecht and Evert, 2009).

2. For corpus-based variational linguistics, corpora of CMC discourse are an important resource that closes the "CMC gap" in corpora of contemporary written language and language-in-interaction. With a considerable part of contemporary everyday communication being mediated through CMC technologies, up-to-date investigations of language change and linguistic variation need to be able to include CMC discourse in their empirical analyses.

In order to harness the full potential of corpus-based research, the preparation of any type of linguistic corpus which includes CMC discourse—whether a genuine CMC corpus or a broad-coverage Web corpus—faces the challenge of handling and annotating the linguistic peculiarities characteristic for the types of written discourse found in CMC genres. Two fundamental (but non-trivial) tasks are (i) accurate tokenization and (ii) sufficiently reliable part-of-speech (PoS) annotation. Together, they provide a layer of basic linguistic information on the token level that is a pre-

Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task, pages 44–56,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

requisite for any form of advanced linguistic analysis on the word, sentence and interaction level.

The linguistic peculiarities of discourse in CMC and social media genres have been extensively described in the literature (for an overview of features with a focus on German CMC see e.g. Haase et al., 1997; Runkehl et al., 1998; Beißwenger, 2000; Storrer, 2001; Dürscheid, 2005; Androutsopoulos, 2007; Bartz et al., 2013; for English CMC see e.g. Crystal, 2001, 2003; Herring, 1996, 2010, 2011). Due to its dialogic nature and depending on the degree to which the interlocutors consider their interaction as an informal, private exchange, CMC discourse typically includes a range of deviations from the syntactic and orthographic norms of the written standard (often referred to as non-canonical phenomena) such as colloquial spellings (e.g., clitics and schwa elisions) and lexical items which typically occur in spoken interactions rather than monologic texts (interjections, intensifiers, focus and gradation particles, modal particles and downtoners, etc.). The word order and syntax of CMC posts exhibit features that are characteristic of spoken or "conceptually oral" language use in colloquial registers (e.g., ellipses, German *weil* or *obwohl* with V2 clause). High speed typing causes speedwriting phenomena such as typos, the omission of upper case or the use of acronyms; other deviations from the orthographic standard have to be considered as intended, creative spellings (*nice2CU*, *good n8*). The need for emotion markers leads to the use of emoticons and emoji; upper case and letter iterations serve as suprasegmental forms of emphasis in the written medium (*LASS DAS!*, *suuuuuper!!!!*). Addressing terms and hashtags indicate reference between user posts and link individual posts to discourse topics.

Tackling the linguistic peculiarities of CMC data with NLP tools is an open issue in corpus and computational linguistics, which has been addressed by an increasing number of papers and approaches over the past years (as a desideratum e.g. Beißwenger and Storrer, 2008; King, 2009; for the development of NLP tools e.g. Ritter et al., 2011; Gimpel et al., 2011; Owoputi et al., 2015; Avontuur et al., 2012; Bartz et al., 2013; Neunerdt et al., 2013; Rehbein, 2013; Rehbein et al., 2013; Horbach et al., 2015; Zinsmeister et al., 2014; Ljubešić et al., 2015). Issues of processing and annotating CMC data have also been a central topic of the DFG-funded scientific network *Empirical Research of Internet-Based Communication* (*Empirikom*), which brought together researchers interested in building and analyzing CMC, social media and Web corpora for research questions in linguistics, computational linguistics and language technology during the years 2010–2014.[1] As a result from discussions in the network, it was decided to set up a community shared task to foster the development of approaches for automatic linguistic annotation of CMC data for German in a competitive setting. The task was named *Empirikom* Shared Task on Automatic Linguistic Annotation of Computer-Mediated Communication and Social Media (EmpiriST 2015).

The design of EmpiriST 2015 was based on the following two premises:

1. It should take into consideration not only the compilation of CMC corpora for research and teaching purposes in linguistics but also the handling of portions of CMC data as part of large Web corpora.
2. It should be based on a freely available gold standard created with a well-defined PoS tagset and precise guidelines for tokenization and PoS annotation (see Sec. 2).

The main goals and research questions are:

1. To what extent can the performance of automatic tools for tokenization and PoS tagging of German CMC discourse be improved, using our gold standard for training or domain adaptation?
2. Can both genuine CMC corpora and Web corpora (where CMC phenomena typically occur much less frequently) be processed by the same approaches and models, or do we need different tools for the two types of corpora?

2 The EmpiriST gold standard

The gold standard developed for the shared task comprises roughly 10,000 tokens of training data provided to participants as well as roughly 10,000 tokens of unseen test data used in the evaluation phase. It was compiled from data samples considered representative for the two types of corpora: (i) a CMC subset covering discourse from a range of CMC/social media genres, and (ii) a Web corpora subset containing CC-licensed Web pages from different genres.

[1] http://www.empirikom.net/

2.1 Data sets

The **CMC subset** includes samples from several CMC genres and different sources:

- a selection of donated tweets from (i) the Twitter channel of an academy project used for (monologic) project-related announcements, (ii) the Twitter channel of a lecturer used for discussions with the students accompanying a university class (= dialogic use of tweets);
- a selection of data taken from the *Dortmund Chat Corpus* (Beißwenger, 2013) representing discourse from different types of chat: (i) *social chat* recorded in multiparty chatrooms where people met mainly for recreational purposes, (ii) *professional chat* comprising professional uses of chatrooms, e.g. advisory chats and chats in the context of learning and teaching;
- a selection of threads retrieved from Wikipedia talk pages;
- a selection of WhatsApp interactions taken from the data collected in the project *Whats up, Deutschland?*;[2]
- a selection of blog comments from CC-licensed weblogs collected by Adrien Barbaresi.

For the **Web corpora subset**, roughly 50,000 running words of text were collected by Web crawling. In order to ensure a broad coverage of Web genres and topics, the crawl was based on a set of manually pre-selected seed words. The following list gives an impression of the distribution of genres in the data:

- Web sites on topics such as hobbies, travel and IT;
- blogs on topics such as hobbies, travel and legal issues;
- Wikipedia articles on topics such as biology, botany and cities;
- Wikinews on topics such as IT security and ecology.

The largest portion of these data is comprised of Web pages, blog entries and commentaries, a smaller portion consists of genres such as Wikipedia articles, Wikinews etc. An important requirement was that all texts must be published under a suitable Creative Commons licence so that the resulting corpus can be made freely available to the community without any legal issues.

From the available data, we selected roughly 5,000 tokens of training data for each subset, which were provided to task participants with manual tokenization and PoS tagging. Another 5,000 tokens per subset were used as unseen test data, with a similar distribution of genres and sources as in the training data. The precise data sizes of the training and test sets are listed in Tab. 1.

	CMC subset	Web subset
training data	5,109 (8 samples)	4,944 (11 samples)
test data	5,234 (6 samples)	7,568 (12 samples)

Table 1: Sizes of the training and test data sets, specified in number of tokens (above) and number of text samples (below).

2.2 Annotation guidelines

For **tokenization**, we developed a guideline with detailed rules for handling CMC-specific tokenization issues (Beißwenger et al., 2015a). It was tested and refined for a range of CMC and Web genres with the help of several student annotators in Berlin, Darmstadt, Dortmund and Erlangen.

For **PoS tagging**, we used the 'STTS_IBK' tag set which had been defined as a result from discussions in the *Empirikom* network and at three workshops dedicated to the adaptation and extension of the canonical version of the *Stuttgart-Tübingen-Tagset* ('STTS 1.0'; Schiller et al., 1999) to the peculiarities of "non-standard" genres (Zinsmeister et al., 2013, 2014). STTS_IBK introduces two types of new tags: (i) tags for phenomena that are specific to CMC and social media discourse, (ii) tags for phenomena that are typical for spontaneous (spoken or "conceptually oral") language in colloquial registers (cf. Tab. 2). These extensions are useful for corpus-based research on CMC as well as spoken conversation. STTS_IBK is downward compatible to STTS 1.0 and therefore allows for interoperability with existing corpora and tools. In addition, the tag set extensions in STTS_IBK are compatible with the STTS extensions defined at IDS Mannheim for the PoS

[2] http://www.whatsup-deutschland.de/

PoS tag	Category	Examples
I. Tags for phenomena specific for CMC / social media discourse:		
EMOASC	ASCII emoticon	`:-) :-( ^^ O.O`
EMOIMG	Graphic emoticon (emoji)	☺ ☹ ☻
AKW	Interaction word	`*lach*, freu, grübel, *lol*`
HST	Hash tag	`Kreta war super! #Urlaub`
ADR	Addressing term	`@lothar: Wie isset so?`
URL	Uniform resource locator	`http://tu-dortmund.de`
EML	E-mail address	`peterklein@web.de`
II. Tags for phenomena typical for spontaneous spoken ('conceptually oral') language in colloquial registers:		
VVPPER	Tags for different types of colloquial contractions thate are frequent in CMC (APPRART already exists in STTS 1.0)	`schreibste, machste`
APPRART		`vorm, überm, fürn`
VMPPER		`willste, darfste, musste`
VAPPER		`haste, biste, isses`
KOUSPPER		`wenns, weils, obse`
PPERPPER		`ichs, dus, ers`
ADVART		`son, sone`
PTKIFG	Intensifier, focus and gradation particles	`sehr schön, höchst eigenartig, nur sie, voll geil`
PTKMA	Modal particles and downtoners	`Das ist ja / vielleicht doof. Ist das denn richtig so? Das war halt echt nicht einfach.`
PTKMWL	Particle as part of a multi-word lexeme	`keine mehr, noch mal, schon wieder`
DM	Discourse markers	`weil, obwohl, nur, also,` ...*with V2 clauses*
ONO	Onomatopoeia	`boing, miau, zisch`

Table 2: Tagset extensions for CMC phenomena in STTS_IBK. More examples with context can be found in the detailed annotation guidelines on the EmpiriST Web site (available in German and English).

annotation of FOLK[3], the Mannheim "Research and Teaching Corpus of Spoken German" (Westpfahl and Schmidt, 2013; Westpfahl, 2013). The tag set is described in an annotation guideline (Beißwenger et al., 2015b) and has been tested with data from several CMC genres in advance.

The complete annotation guidelines (in German) as well as supplementary documentation are available online from the shared task Web site.[4] For international participants, an English translation of the tagging guideline is also provided.

2.3 Annotation procedure

All data sets were manually tokenized and PoS tagged by multiple annotators, based on the official tokenization (Beißwenger et al., 2015a) and tagging guidelines (Schiller et al., 1999; Beißwenger et al., 2015b), see Sec. 2.2. Cases of disagreement were then adjudicated by the task organizers to produce the final gold standard. During the annotation of the training data, minor changes to the annotation guidelines were made based on experience from the adjudication procedure. In addition, various problematic cases were collected in a supplementary document available to the annotators.

The manual tokenization was carried out in a plain text editor, starting from whitespace-tokenized files in one-token-per-line format. Annotators were instructed to make no other changes to the files than inserting additional line breaks as token boundaries (except for a few special cases), but were allowed to mark unclear cases with comments. The tokenizations were compared and adjudicated using the `kdiff3` utility.[5]

In the next step, manual tagging was partly carried out with the Web-based annotation platform *CorA*[6] (Bollmann et al., 2014), partly with

[3] `http://agd.ids-mannheim.de/folk.shtml`

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

[5] `http://kdiff3.sourceforge.net/`

[6] `https://www.linguistics.rub.de/comphist/resources/cora/`

	BT	FW
gold	96.04	94.05
BT		91.05

Table 3: Agreement between annotators and gold standard for PoS tagging of the CMC data subset (training and test sets). Values are accuracy (acc) percentages.

our own Web-based tool *MiniMarker*. In both cases annotators worked independently with separate password-protected accounts and were encouraged to document interesting or difficult phenomena in free-form comments. CorA has the advantage that tokenization errors can be corrected at the tagging stage, while MiniMarker enables annotators to look up how specific word forms are tagged in the TIGER treebank corpus in order to ensure consistent annotation. For adjudication of the PoS tagging, we pre-annotated unanmimous annotator decisions and filled in the remaining disputed tags with MiniMarker.

Agreement between annotators as well as the agreement of each annotator with the final gold standard was determined using the same evaluation metrics as for systems participating in the shared task (see Sec. 3.2).

2.3.1 CMC subset

In a preliminary study on the manual tokenization of CMC (cf. Beißwenger et al., 2013), we observed very high inter-annotator agreement with F_1 scores ranging from 98.6% to 99.7%, showing that manual tokenization of such data provides a valid and reliable gold standard. For training and test data of the CMC subset, we therefore decided to pursue a "sequential double keying" approach. The initial tokenization was done at a very early stage of the task preparation; it was later double-checked and revised according to the final tokenization guidelines by a second expert annotator.

PoS tags were added by two independent annotators. Tab. 3 shows the observed agreement between the annotators and the adjudicated gold standard in terms of accuracy (acc).

Frequent errors involved the new particle classes in STTS_IBK (PTKIFG, PTKMA, PTKMWL), punctuation (`$(` vs. `$.`), the distinction between common (NN) and proper nouns (NE) and the correct classification of non-inflected adjectives (ADJD).

It is interesting to note that for both annotators the agreement between each annotator and the gold standard is much higher than the agreement between the two annotators. One possible explanation is that each annotator had difficulties with specific types of phenomena. Looking at the error classes, this assumption turns out to be true: For example, annotator FW tended to misclassify adverbs as intensifier particles (PTKIFG, $n = 66$) whereas annotator BT made this mistake only six times. On the other hand, BT misjudged more than twice as many adjectives (ADJA vs. ADJD) than FW.

2.3.2 Web corpora subset

The test data of the Web corpora subset were manually tokenized by five primary annotators, and then adjudicated in two phases by one of the task organizers. Tab. 4 shows pairwise agreement between annotators and the agreement of each annotator with the gold standard in terms of F_1 scores for token boundaries. Agreement is very high between all pairs of annotators, indicating that the manual tokenization is reliable.

	AM	AS	DP	JM	LS
gold	99.56	99.74	99.70	99.78	99.93
AM		99.75	99.67	99.66	99.62
AS			99.88	99.89	99.80
DP				99.87	99.71
JM					99.73

Table 4: Agreement between annotators and gold standard for tokenization of the Web corpora test data. Values are F_1 scores given as percentages.

	AM	AS	JM	LS
gold	92.64	96.15	95.49	91.77
AM		91.54	90.80	88.42
AS			93.04	89.51
JM				90.27

Table 5: Agreement between annotators and gold standard for PoS tagging of the Web corpora test data. Values are accuracy (acc) percentages.

PoS tags were manually added by 4 independent annotators, based on the adjudicated tokenization. No further corrections of the tokenization were found to be necessary in this phase. Tab. 5 shows agreement between the annotators and the gold standard in terms of observed accuracy (acc). Due to the low probability of chance agreement (approx. 7.5%), there is no need to compute κ values or other adjusted scores. Agreement for the manual tagging is less satisfactory than for the tokenization. Major sources of disagreement were the newly introduced particle classes—in particular PTKIFG and PTKMA—as well as unintuitive or poorly defined category boundaries in the original STTS 1.0 tag set—in particular common nouns (NN) vs. proper nouns (NE) vs. foreign text (FM), and adverbs (ADV) vs. adverbial adjectives (ADJD). It is also noticeable that the training and experience of individual annotators played an important role: two annotators (AS and JM) agree fairly well with each other and with the adjudicated gold standard, while the other two annotators performed considerably worse.

Despite these issues, most errors and misinterpretations were caught by our adjudication of the four-fold annotation. A fifth independent tagging carried out by annotator SM at a later stage showed an agreement of acc $= 95.90\%$ with the final gold standard.

The training data of the Web corpora subset were manually tokenized by three independent annotators and tagged by five independent annotators, with adjudication by one of the task organizers after each stage. Agreement between annotators and the gold standard is similar to the test data.

2.4 Availability

All gold standard data sets, the specification of the extended STTS tag set and the guidelines for tokenization and PoS tagging have been published on the EmpiriST Web site[7] and will remain available for use in future research. We used simple UTF-8 encoded text formats for both raw and annotated versions of the data. Annotated files are provided in one-token-per-line format with empty lines serving as posting or paragraph boundary markers. Corresponding PoS tags are given in an additional column separated from the token text by a single tab stop. Metadata for each posting or Web page are inserted as empty XML elements on separate lines. A small excerpt from one of the files is shown in Fig. 1.

Apart from the actual contents, the EmpiriST 2015 data package comes with a description of the tag set, evaluation scripts and licensing information. All files are released under the *Creative Commons* CC BY-SA 3.0 licence.[8]

```
<posting info="User 15:08, 26.09.10" />
Das     ART
ständige        ADJA
Revertieren     NN
von     APPR
Phi     NE
damit   PAV
auch    ADV
...     $.
```

Figure 1: Excerpt from the CMC subset of the EmpiriST 2015 shared task training data.

3 The shared task

3.1 Layout of the task

The EmpiriST 2015 shared task was divided into three major stages: (i) preparation, (ii) training and (iii) evaluation.

The preparation stage started with the release of the annotation guidelines together with roughly 2,000 tokens of trial data from each subset in October 2015. The trial data were intended to illustrate the required input and output file formats and to give an impression of the specific characteristics of the CMC and Web texts to be processed. They were based on preliminary versions of the guidelines and were produced without multiple annotation. Participants were instructed that they should not be relied on for training the final systems. During the preparation stage, there was also a fruitful dialogue between interested parties and the shared task organizers, leading to clarifications and corrections of the guidelines.

The second stage was dedicated to the training and adaptation of the competing systems. It started with the release of the complete training data on the shared task Web site in December 2015. The registration deadline fell within this stage, enabling participants to make an initial assessment of their performance before registering.

The evaluation stage was divided into two consecutive phases so that (i) tokenization and tagging quality could be evaluated separately and (ii) the same test data could be used for both subtasks. In each phase, unannotated test data were released via the shared task Web site; participants then had to submit their system output within five days by e-mail. For the tokenization phase, raw texts were

[7] https://sites.google.com/site/empirist2015/home/gold

[8] https://creativecommons.org/licenses/by-sa/3.0/

released, padded with additional filler data in order to prevent tuning of systems to the test data before the second phase. For the tagging phase, manually tokenized versions of the texts were released. The two phases took place in two consecutive weeks in February 2016.

3.2 Evaluation metrics

Evaluation of the submissions to EmpiriST 2015 was carried out by the task organizers. Following Jurish and Würzner (2013), results for the tokenization task were evaluated based on the unweighted harmonic average (F_1) between precision (pr) and recall (rc) of the token boundaries in the participants' submissions. Formally, let $B_{\text{retrieved}}$ be the set of token boundaries predicted by the tokenization procedure to be evaluated and B_{relevant} those present in the gold standard; then:

$$\mathrm{pr} = \frac{|B_{\text{relevant}} \cap B_{\text{retrieved}}|}{|B_{\text{retrieved}}|} \quad (1)$$

$$\mathrm{rc} = \frac{|B_{\text{relevant}} \cap B_{\text{retrieved}}|}{|B_{\text{relevant}}|} \quad (2)$$

$$\mathrm{F}_1 = \frac{2 \cdot \mathrm{pr} \cdot \mathrm{rc}}{\mathrm{pr} + \mathrm{rc}} \quad (3)$$

For technical reasons, the trivial token boundary at the beginning of each text file is included in the evaluation, but not the boundary at its end.[9]

Following Giesbrecht and Evert (2009), the PoS tagging task was evaluated in terms of the accuracy (acc) of the PoS tag assignments in the participants' submissions. Formally, let n_{correct} be the number of tokens whose tags agree with the gold standard, and n_{total} the total number of tokens in the data set; then:

$$\mathrm{acc} = \frac{n_{\text{correct}}}{n_{\text{total}}} \quad (4)$$

In order to support participants in development and self-evaluation of their submissions, both evaluation metrics were implemented as Perl scripts by the organizers and published together with the training and test data sets.

4 Participating systems

Tab. 6 gives an overview of the participating teams and systems. Team UdS submitted three related systems (UdS-distributional, UDS-retrain, UDS-surface). In addition, each system was permitted to submit up to 3 different runs, with only the best run being included in the task results.

Team	Reference
	Tokenization
AIPHES	Remus et al. (2016)
COW	Schäfer and Bildhauer (2012)[1]
LTL-UDE	Horsmann and Zesch (2016)
SoMaJo	Proisl and Uhrig (2016)
$WAGMOB[†]	—
	PoS tagging
AIPHES	Remus et al. (2016)
bot.zen[*]	Stemle (2016)
COW[†]	Schäfer and Bildhauer (2012)[1]
LTL-UDE	Horsmann and Zesch (2016)
$WAGMOB[†]	—
UdS	Prange et al. (2016)

[*] late submission
[†] non-competitive submission
[1] see also Schäfer (2015)

Table 6: Overview of the participants with reference to the corresponding system description.

4.1 Summary of competing approaches

As shown in Tab. 6, we had five submissions for the **tokenization** subtask, one of them non-competitive.[10] All five systems employed rule-based tokenization approaches. Two of them (AIPHES and LTL-UDE) used a "split and merge" strategy that splits tokens into atomic units in the first pass. In subsequent passes, higher-order rules implement merging strategies for dealing with complex phenomena such as URLs, abbreviations or emoticons. In contrast, COW used an "under segmentation" strategy protecting certain token sequences in the first pass and further segmenting them in a second. SoMaJo used complex, cascaded regular expressions successively dealing with the aforementioned classes of phenomena.

All approaches made use of additional lists of abbreviations, proper names, emoticons, etc. in order to improve correct tokenization of special characters and punctuation.

We had six submissions for the **PoS tagging** subtask, two of them non-competitive.[11] From the

[9] This trick simplified the implementation of the evaluation script considerably. It was deemed to be acceptable with a typical effect of less than 0.01% on the evaluation metrics.

[10] $WAGMOB was a student team from a Bachelor seminar taught by one of the task organizers

[11] COW is an existing annotation pipeline for large Web corpora, which was entered into the task with minimal adap-

four regular submissions, one (bot.zen) was sent in after the submission deadline and is thus not included in the official ranking. In contrast to tokenization, all systems competing in the PoS tagging subtask made use of statistical models specially trained or re-trained for the purpose of EmpiriST 2015. The types of models employed reflect all state-of-the-art approaches to the task of PoS tagging. All approaches have in common that they extend the EmpiriST training data with additional corpora and linguistic resources.

The three UdS systems built on a classical *hidden Markov model* (HMM; Rabiner, 1989). In addition, they focused on improvements in the analysis of out-of-vocabulary (OOV) words by adding domain-specific training material and a list of likely PoS tags for OOV items. LTL-UDE and AIPHES used *conditional random fields* (CRF; Lafferty et al., 2001). Both systems differed in the selection of features and the additional resources used in the training process. Team bot.zen employed a long short-term memory (LSTM; Hochreiter and Schmidhuber, 1997) recurrent neural network in combination with neural word embeddings as input representations (Mikolov et al., 2013).

5 Results

In order to put the performance of the shared task submissions into perspective, we also evaluated several widely-used off-the-shelf tools as baselines:

- the WASTE tokenizer (Jurish and Würzner, 2013);[12]
- TreeTagger v3.2 (Schmid, 1995);[13]
- Stanford tagger v3.6.0 (Toutanova et al., 2003);[14]
- the COW pipeline (Schäfer and Bildhauer, 2012; Schäfer, 2015).[15]

Tab. 7 (tokenization) and Tab. 8 (PoS tagging) show the results obtained by all task participants and baseline systems on the CMC and Web corpora subsets. Within each subset, results are micro-averaged across the text samples. The overall score is the macro-average over both subsets, ensuring that CMC and Web corpora carry the same weight. For systems that submitted multiple runs, only the best run is shown in the table (indicated by a subscript appended to the team name). The official ranking ("podium") includes only competitive and timely submissions. Since team UdS entered three closely related systems into the competition, only one of them was selected for the official podium. Detailed results for individual runs and text samples are available on the EmpiriST Web page.[16]

Since the existing off-the-shelf taggers used as a baseline are not aware of the new PoS tags in STTS_IBK, the evaluation was carried out both at the level of STTS_IBK and at the level of the established STTS 1.0 tag set (Schiller et al., 1999). For this purpose, one or more alternative STTS 1.0 tags were also accepted for each extended tag in the gold standard. The precise mapping rules are specified in Tab. 9. The official ranking is always based on the full STTS_IBK tag set.

6 Conclusion

The systems submitted to the EmpiriST2015 shared task have improved the state-of-the-art for tokenization and PoS tagging of CMC and Web corpora. The best submitted tokenizer achieved an F_1-score of 99.54% (vs. 98.47% baseline) for the CMC data set and an F_1-score of 99.77% (vs. 99.42% baseline) for the Web corpora data set. For PoS tagging, the results are still far from optimal. Nevertheless, the improvement against baseline systems is striking especially for the CMC subset: The best submitted tagger achieved an accuracy of 87.33% evaluated against STTS_IBK (vs. 77.89% baseline), and an accuracy of 90.28% against STTS 1.0 (vs. 81.51% baseline). For the Web corpora subset, where the baseline systems already peform much better than on gen-

tations to account for the tokenization principles and extended tag set of EmpiriST. It may therefore be more appropriate to compare COW with the baseline systems than with the other task participants.

[12] We used WASTE as shipped with the moot package (v2.0.13, `http://kaskade.dwds.de/waste/`) and trained a model solely using the EmpiriST training data.

[13] We used the German UTF-8 parameter file downloaded from `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/` on 21 June 2016.

[14] We used the `german-dewac` parameter file from the distribution released on 9 Dec 2015. Substantial automatic and manual post-editing was required to undo character transformations made by the tokenizer, replace non-STTS tags (e.g. `$[` instead of `$(`), and account for the systematic mistagging of parentheses and brackets as `TRUNC`.

[15] COW results were submitted by the developers as a baseline participation in the PoS tagging subtask.

[16] `https://sites.google.com/site/empirist2015/home/results`

	CMC				Web				Overall		
Team	pr	rc	F$_1$	Rk	pr	rc	F$_1$	Rk	F$_1$	Rk	**Pdm**
SoMaJo	99.52	99.56	99.54	1	99.57	99.64	99.60	3	99.57	1	**1**
AIPHES	99.30	98.62	98.96	2	99.63	99.89	99.76	2	99.36	2	**2**
COW	98.31	98.07	98.18	5	99.84	99.71	99.77	1	98.98	3	**3**
WASTE†	99.41	97.57	98.47	4	99.59	99.26	99.42	4	98.95	4	-
LTL-UDE	99.01	98.18	98.58	3	98.92	99.54	99.22	8	98.90	5	**4**
\$WAGMOB*	98.97	96.79	97.83	6	99.41	99.38	99.39	5	98.61	6	-
Stanford†	97.19	97.69	97.41	7	98.97	99.71	99.34	7	98.38	7	-
TreeTagger†	94.95	95.01	94.96	8	99.58	99.14	99.36	6	97.16	8	-

Table 7: Results of the tokenization subtask including non-competitive submissions (marked with *) and baseline systems (marked with †). The last column gives the official EmpiriST 2015 “podium” ranking. pr, rc, and F$_1$ are given as percentages for better readability.

	CMC				Web				Overall		
	STTS_IBK		STTS 1.0		STTS_IBK		STTS 1.0		STTS_IBK		
Team	acc	Rk	acc	Rk	acc	Rk	acc	Rk	acc	Rk	**Pdm**
UdS-distributional$_2$	87.33	1	90.28	1	93.55	1	94.62	1	90.44	1	**1**
UdS-retrain$_2$	86.40	3	89.07	3	92.79	3	93.86	3	89.60	2	-
UdS-surface$_2$	86.45	2	89.28	2	92.43	4	93.50	4	89.44	3	-
LTL-UDE$_2$	86.07	4	88.84	4	92.10	5	93.12	5	89.09	4	**2**
AIPHES	84.22	7	87.10	6	93.27	2	94.30	2	88.75	5	**3**
bot.zen$_3^*$	85.42	5	87.47	5	90.63	8	91.74	9	88.03	6	-
COW†	77.89	8	81.51	8	91.82	6	92.96	6	84.86	7	-
\$WAGMOB*	84.77	6	87.03	7	84.51	10	85.57	10	84.64	8	-
TreeTagger†	73.21	9	76.81	9	91.75	7	92.89	7	82.48	9	-
Stanford†	70.60	10	75.83	10	89.42	9	92.52	8	80.01	10	-

Table 8: Results of the PoS tagging subtask including non-competitive or late submissions (marked with *) and baseline systems (marked with †). If applicable, a subscript indicates the best run of the respective system (based on overall accuracy), which is listed in the table. The last column gives the official EmpiriST 2015 “podium” ranking. acc is given as a percentage for better readability.

gold tag	these tags are also accepted
EMOIMG	XY ITJ EMOASC
AKW	VVFIN VVIMP VVINF VVIZU VAFIN VAIMP VAINF VMFIN VMINF
HST	XY
ADR	XY NE
URL	XY
EML	XY
VVPPER	VVFIN VVIMP VVINF
VMPPER	VMFIN VMINF
VAPPER	VAFIN VAIMP VAINF
KOUSPPER	KOUS
PPERPPER	PPER
ADVART	ART
PTKIFG	ADV ADJD PTKMA PTKMWL
PTKMA	ADV ADJD PTKIFG PTKMWL
PTKMWL	ADV ADJD PTKIFG PTKMA
DM	KOUS ADV
ONO	ITJ VVFIN VVIMP VVINF
ADV	PTKIFG PTKMA PTKMWL DM
KOUS	DM
PIDAT	PIAT

Table 9: Mapping of extended tags for evaluation at the level of STTS 1.0.

uine CMC, there was only a modest improvement: 93.55% against STTS_IBK (vs. 91.82% baseline), and 94.62% against STTS 1.0 (vs. 92.96% baseline). It should be noted that the widely-used Stanford and TreeTagger tools performed substantially worse on tagging CMC than the COW baseline shown here.

Further evaluation of the results in future work should include a close examination and discussion of the performance of the tagger models with respect to the tag set extensions defined in STTS_IBK, as well as their performance on different genres and text sources. This will be the topic of a round table organized at the 3rd NLP4CMC workshop at KONVENS 2016.[17]

The results of the shared task can be considered a promising step towards better NLP tools for German CMC data, especially since all participants (except for UdS) have made their systems available to the community as open-source software. However, the adaptation of NLP tools to the linguistic peculiarities of CMC discourse—especially for PoS tagging—is still a challenging task. The resources developed for EmpiriST 2015 (gold standard and annotation guidelines) will remain available on the task Web site under a Creative Commons licence.[18] We hope that they will stimulate further advances in adapting NLP technologies to CMC discourse as well as in improving the annotation quality of German Web corpora.

Acknowledgments

We thank the *German Society for Computational Linguistics and Language Technology* (*GSCL*) for financial support, Stefanie Dipper and Marcel Bollmann (Ruhr-Universität Bochum) for providing the *CorA* tool for PoS annotations, our student annotators in Berlin, Darmstadt, Dortmund and Erlangen for their careful work and the members of the DFG network *Empirikom* for fruitful discussions in the design stage of the task.

References

Jannis Androutsopoulos. 2007. Neue Medien – neue Schriftlichkeit? *Mitteilungen des Deutschen Germanistenverbandes* 54(1):72–97.

Tetske Avontuur, Iris Balemans, Laura Elshof, Nanne van Noord, and Menno van Zaanen. 2012. Developing a part-of-speech tagger for Dutch tweets. *Computational Linguistics in the Netherlands Journal* 2:34–51.

Thomas Bartz, Michael Beißwenger, and Angelika Storrer. 2013. Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *Journal for Language Technology and Computational Linguistics (JLCL)* 28(1):157–198.

Michael Beißwenger. 2000. *Kommunikation in virtuellen Welten: Sprache, Text und Wirklichkeit*. Ibidem-Verlag.

Michael Beißwenger. 2013. Das Dortmunder Chat-Korpus. *Zeitschrift für Germanistische Linguistik* 41(1):161–164.

Michael Beißwenger, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. 2013. Preparing a shared task on linguistic annotation of computer-mediated communication. Talk and poster presentation at the International Conference of the GSCL.

Michael Beißwenger, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. 2015a. Richtlinie für die manuelle Tokenisierung von

[17] `https://sites.google.com/site/nlp4cmc2016/`

[18] `https://sites.google.com/site/empirist2015/`

Sprachdaten aus Genres internetbasierter Kommunikation. Empirist2015 Guideline document.

Michael Beißwenger, Thomas Bartz, Angelika Storrer, and Swantje Westpfahl. 2015b. Tagset und Richtlinie für das Part-of-Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation. EmpiriST2015 Guideline document.

Michael Beißwenger and Angelika Storrer. 2008. Corpora of computer-mediated communication. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, Walter de Gruyter, Berlin and New York, volume 1 of *Handbücher zur Sprache und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science*, chapter Corpora of Computer-Mediated Communication, pages 292–308.

Marcel Bollmann, Florian Petran, Stefanie Dipper, and Julia Krasselt. 2014. CorA: A web-based annotation tool for historical and other nonstandard language data. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Gothenburg, Sweden, pages 86–90.

David Crystal. 2001. *Language and the Internet*. CUP, Cambridge.

David Crystal. 2003. *English as a Global Language*. Cambridge University Press, second edition. Cambridge Books Online.

Christa Dürscheid. 2005. Normabweichendes Schreiben als Mittel zum Zweck. *Muttersprache: Vierteljahresschrift für deutsche Sprache / Gesellschaft für Deutsche Sprache (GfdS)* 115(1):40–53.

Eugenie Giesbrecht and Stefan Evert. 2009. Part-of-speech tagging – a solved task? An evaluation of POS taggers for the web as corpus. In Inaki Alegria, Igor Leturia, and Serge Sharoff, editors, *Proceedings of the Fifth Web as Corpus Workshop (WAC5), San Sebastián, Spain, 7 September, 2009*. pages 27–35.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '11, pages 42–47.

Martin Haase, Michael Huber, Alexander Krumeich, and Georg Rehm. 1997. *Internetkommunikation und Sprachwandel*, VS Verlag für Sozialwissenschaften, Wiesbaden, pages 51–85.

Susan C. Herring, editor. 1996. *Computer-Mediated Communication. Linguistic, Social and Cross-Cultural Perspectives*. Pragmatics and Beyond New Series 39. John Benjamins, Amsterdam and Philadelphia.

Susan C. Herring. 2010. Computer-mediated conversation part i: Introduction and overview. *Language@Internet* 7(2).

Susan C. Herring. 2011. Computer-mediated conversation part ii: Introduction and overview. *Language@Internet* 8(2).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–1780.

Andrea Horbach, Stefan Thater, Diana Steffen, M. Peter Fischer, Andreas Witt, and Manfred Pinkal. 2015. Internet corpora: A challenge for linguistic processing. *Datenbank-Spektrum* 15(1):41–47.

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

Bryan Jurish and Kay-Michael Würzner. 2013. Word and Sentence Tokenization with Hidden Markov Models. *JLCL* 28(2):61–83.

Brian King. 2009. *Building and Analysing Corpora of Computer-Mediated Communication*, Continuum, London, volume Contemporary corpus linguistics, pages 301–320.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML '01, pages 282–289.

Nikola Ljubešić, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak, and Iza Škrjanec. 2015. Predicting the level of text standardness in user-generated content. In *Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria, Sep 7–9 2015*. page 371–378.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Melanie Neunerdt, Bianka Trevisan, Michael Reyer, and Rudolf Mathar. 2013. Part-of-speech tagging for social media texts. In *Proceedings of the 25th Conference of the German Society for Computational Linguistics (GSCL 2013)*. pages 139–150.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, and Nathan Schneider. 2015. Part-of-speech tagging for Twitter: Word clusters and other advances. Technical report, Technical report, Carnegie Mellon University (CMU-ML-12-107).

Jakob Prange, Andrea Horbach, and Stefan Thater. 2016. UdS-(retrain|distributional|surface): Improving POS tagging for OOV words in German CMC and web data. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*. Association for Computational Linguistics, Berlin, pages 97–105.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*. Association for Computational Linguistics, Berlin, pages 91–96.

Lawrence R. Rabiner. 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2):257–286.

Ines Rehbein. 2013. Fine-grained pos tagging of German tweets. In *International Conference of the German Society for Computational Linguistics and Language Technology (GSCL2013), September 25-27, Darmstadt, Germany*.

Ines Rehbein, Emiel Visser, and Nadine Lestmann. 2013. Discussing best practices for the annotation of Twitter microtext. In *Proceedings of The Third Workshop on Annotation of Corpora for Research in the Humanities (ACRH-3)*. pages 73–84.

Steffen Remus, Gerold Hintz, Darina Benikova, Thomas Arnold, Judith Eckle-Kohler, Christian M. Meyer, Margot Mieskes, and Chris Biemann. 2016. EmpiriST: AIPHES – robust tokenization and POS-tagging for different genres. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*. Association for Computational Linguistics, Berlin, pages 106–114.

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

Jens Runkehl, Peter Schlobinski, and Torsten Siever. 1998. *Sprache und Kommunikation im Internet : Überblick und Analysen*. Westdt. Verl., Opladen [u.a.].

Roland Schäfer. 2015. 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)*. UCREL, Lancaster, UK.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC '12)*. ELRA, Istanbul, Turkey, pages 486–493.

Anne Schiller, Simone Teufel, and Christine Stöckert. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). University of Stuttgart: Institut für maschinelle Sprachverarbeitung.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT Workshop*.

Egon Stemle. 2016. bot.zen @ EmpiriST 2015 – a minimally-deep learning PoS-tagger (trained for German CMC and web data). In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*. Association for Computational Linguistics, Berlin, pages 115–119.

Angelika Storrer. 2001. *Getippte Gespräche oder dialogische Texte? Zur kommunikationstheoretischen Einordnung der Chat-Kommunikation*, Walter de Gruyter, Berlin, New York, volume Sprache im Alltag. Beiträge zu neuen Perspektiven in der Linguistik. Herbert Ernst Wiegand zum 65. Geburtstag gewidmet, page 439–465.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Swantje Westpfahl. 2013. STTS 2.0? Improving the tagset for the part-of-speech-tagging of German spoken data. In Manfred Levin, Lori und Stede, editor, *Proceedings of LAW VIII – The 8th Linguistic Annotation Workshop. Dublin, Ireland: Association for Computational Linguistics and Dublin City University*. Association for Computational Linguistics and Dublin City University, pages 1–10.

Swantje Westpfahl and Thomas Schmidt. 2013. Pos für(s) FOLK – Part of Speech Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch. *Journal for Language Technology and Computational Linguistics* 28(1):139–153.

Heike Zinsmeister, Ulrich Heid, and Kathrin Beck. 2013. Das Stuttgart-Tübingen Wortarten-Tagset - Stand und Perspektiven. *Special Journal for Language Technology and Computational Linguistics* 28(1).

Heike Zinsmeister, Ulrich Heid, and Kathrin Beck. 2014. Adapting a part-of-speech tagset to non-standard text: The case of STTS. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. LREC 2014, pages 4097–4104.