

Part-of-speech tagging for corpora of computer-mediated communication: A case study on finding rare phenomena

Michael Beißwenger, University of Duisburg-Essen

Tobias Horsmann, University of Duisburg-Essen

Torsten Zesch, University of Duisburg-Essen

Abstract

The paper reports on experiments in the adaptation of part-of-speech (PoS) tagging technology for written, interactional discourse retrieved from social media genres (*computer-mediated communication, CMC*). Starting from an overview of related approaches, we give a summary of the results and discuss lessons learned from a community shared task on PoS tagging German CMC conducted in 2016. These results suggest that further effort should be put into the development of solutions for phenomena which, on the one hand, are too sparsely represented in data samples that could be used for training tagger models, but, on the other hand, are of special interest for the annotation of linguistic corpora. We present a case study in which we used a PoS tagger to find one particular phenomenon of that type, namely German verb-pronoun contractions, in chats and tweets. Whereas the adoption of over- and undersampling strategies to artificially enhance the frequency of the phenomenon in the training data does not lead to significant improvements, the choice of the tagger together with the expansion of the training data with relatively small amounts of additionally labelled instances turns out to be a promising way to let the tagger learn the local word context, and thus improve the recall of the phenomenon in focus while sustaining a high level of precision.

Keywords: CMC, social media, NLP, annotation, PoS tagging

1 INTRODUCTION

This paper reports on experiments in the adaptation of part-of-speech (PoS) tagging technology for written, interactional discourse retrieved from social media environments (tweets, chats, forums, blogs, wikis, social network sites, SMS, WhatsApp, Instagram, etc.). We refer to this type of written, interactional discourse as *computer-mediated communication* (CMC) and to the environments where CMC can be found (be it exclusively, as in the case of chatrooms, or as one among other types of discourse, as on Facebook and Wikipedia) as *social media*. The main challenge of adapting natural language processing (NLP) tools for an accurate automatic annotation of CMC data is dealing with linguistic peculiarities which result (i) from the dialogic, interactional conception of the written utterances, and (ii) from a spontaneous production strategy which is commonly adopted by CMC users, especially in informal settings. Starting from an overview of approaches that have been developed to deal with this issue (Section 2), and from an outline of the views of language technologists and linguists on PoS tagging of CMC data (Section 3), we give a summary of the results and discuss lessons learned from a community shared task on PoS tagging German CMC (*EmpiriST*) conducted in 2016 (Section 4). These results suggest that more effort should be put into the development of solutions for dealing with phenomena which, on the one hand, are too sparsely represented in data samples that could be used for training tagger models, but, on the other hand, are of special interest for the annotation of linguistic corpora. In Section 4, we present a case study in which we used a PoS tagger to find one particular phenomenon of this type, namely German verb-pronoun contractions (*haste, schreibste, gibts, geht's, ...*) in chats and tweets. The results open up some directions for further work and suggest close cooperation between language technologists and linguists as a promising approach for further advances in the automatic identification of rare phenomena in corpora.

2 STATE-OF-THE-ART

Robust part-of-speech (PoS) tagging of CMC still poses a challenge. Instead of tagging accuracy in the high nineties, as on edited text, which is close to the written standard (as can be found in newswire texts and similar text types), we see a big performance drop on CMC, where we only get accuracies of around 80% (Ritter et al. 2011) or even less, depending on the genre (e.g., 69% as a baseline for German chats, as reported by Horbach et al. 2014). The main reason for this performance drop, as noted in Eisenstein (2013), is the high number of out-of-vocabulary words in CMC. Authors, for instance, may neglect orthographic rules

and join, add, omit, or swap letters. Bartz et al. (2013) give a typology of linguistic phenomena which affect this performance, and group them into six main types (with subtypes): speedwriting phenomena, written emulations of prosody, colloquial spellings, creative spellings, CMC-specific acronyms and CMC-specific ‘interactive units,’ which include emoticons, addressing terms and German inflectives. The dialogic character of written utterances in CMC, moreover, also affects syntax, as for example personal pronouns at the beginning of sentences are often omitted (*ellipsis*), as in “went to the gym,” where the pronoun ‘I’ is implied (Ritter et al. 2011). There are two main paradigms to tackle these challenges, normalisation and domain adaptation, as discussed below.

Normalisation removes the orthographic and syntactical anomalies of a text and brings them into their correct form (Han and Baldwin 2011, Chrupala 2014). The text is fitted to the tagger, which is usually trained on edited text, prototypically newswire text, which enables the tagger to perform well. Easy as this might sound, normalisation is probably a more challenging task than domain adaptation. In order to perform normalisation, one has to know (i) that a certain word form is a non-standard form, and (ii) how to normalise it. This entails two tasks, detection and correction. For both steps, an external knowledge source is needed which, especially for the CMC domain, with its many non-standard word forms, can be expected to have a coverage problem. Since performance depends on the degree of coverage obtained, the resulting normalised sequence is not necessarily easier to tag. As such, we will use the second paradigm, domain adaptation, which is more suited to the current work, since it operates directly on the word forms as they appear in CMC data.

Domain Adaptation uses PoS annotated text from the CMC domain to retrain the tagger. The tagger thus learns the characteristics of the domain and is then able to tag CMC data with high accuracy. As existing manually annotated CMC data sets are rather small, a strategy to compensate for this data sparsity problem is to add knowledge from other discourse domains. There are two main strategies for this described in the literature. First, to add more labelled training data by adding foreign domain or machine-generated data (Daumé III 2007; Ritter et al. 2011). Machine-generated data can be created, for instance, by applying several newswire-trained PoS taggers to CMC discourse and adding the related data to the training set when the taggers agree. A second approach is to incorporate external knowledge from resources containing word distributional knowledge, and to guide the machine learning algorithm to extract more information from the existing data (Ritter et al. 2011, Owoputi et al. 2013). The first strategy is related to *which* kind of data is learned, while second to *what* is learned.

The main challenge in tagging CMC lies in dealing with the large number of unknown word forms. Van Halteren and Oostdijk (2014) estimate a range of 20%

to 36% non-word tokens and 4% to 11% out-of-vocabulary (OOV) tokens in (Dutch) tweets. The PoS annotated data sets from the CMC domain are usually too small to cover the high number of word forms which can occur in CMC data, and so cannot yield robust models. While for some languages (e.g. English and German) several data sets exist, these are not easy to combine as the annotation schemes and tagsets used differ, and cannot be easily harmonised.

In the face of these problems with regard to a lack of training data, three methods have been shown to yield considerable improvements with regard to tagging CMC data for English (Ritter et al. 2011, Owoputi et al. 2013) and German (Rehbein 2013, Neunerdt et al. 2013):

1. adding foreign domain data to add lexical and contextual knowledge,
2. adding PoS dictionaries created from other existing corpora,
3. adding word distributional knowledge obtained from unsupervised machine learning methods trained on large collections of plain text.

(1) With the use of *foreign domain data*, text from other existing corpora which have an at least partly compatible PoS tagset is added. Most of the time newswire corpora with edited text are used for this, and these are available for many languages; however, similar-domain text data – such as chat corpora, in the context of the current study – are used if available. Adding more data to the tagger and thus providing more lexical knowledge can be useful in the CMC domain, as it is very useful to know which words can occur together and which inflections are possible for a word (even if only in standard language).

(2) *PoS dictionaries* contain the most frequent PoS tags a word form can have. These dictionaries are created from various corpora, and mainly serve to provide a bias for OOV words. The usefulness of a dictionary is determined by the similarity of the source corpus to the CMC domain and its size. For instance, Neunerdt et al. (2014) created a verb lexicon from a website which also lists common contracted forms that may occur in informal written communication.

(3) *Word distributional knowledge* is provided by applying clustering methods to a large amount of unlabelled data from the CMC domain. Words are clustered according to their distributional similarity, i.e. by a similar word context in which they tend to appear. This property is particularly valuable for PoS tagging of CMC data, as many spelling variations of the same word (e.g. *tomorrow*, *tmr*, *2mr*, *tmrrow*, etc.) tend to be placed into the same cluster (Ritter et al. 2011). If at least one of the word forms in a cluster did occur in the training data, i.e. the correctly spelled form, the tagger receives a bias to assign an unknown word the same tag as that of the known word if both words appear in the same cluster.

The obtained word clusters are identified by ID numbers which can be understood as a kind of PoS tag. According to the similarity function used for clustering, all words which are placed into the same cluster occur in similar word contexts. Hence, one will find clusters with gerund verbs, happy emoticons, sad emoticons, plural nouns, and so on. The number of created clusters usually exceeds the number of tags in human-defined tagsets. Furthermore, the numbering of the clusters is arbitrary, and each time the clustering algorithm is executed the clusters will have different IDs. This arbitrary numbering limits the use of clustering methods for linguists, as cluster IDs are always changing. By using clusters in supervised machine learning scenarios, a mapping from the arbitrary numbering to the tags in a human-defined tagset can be learned, which enables the use of unsupervised methods in supervised setups.

Word clusters have been reported as highly effective if the clustering is applied over a large collection of plain text (Ritter et al. 2011, Rehbein 2013), with Brown clusters (Brown et al. 1992) being frequently used in the literature. Words in Brown clusters are identified by a binary string, and this can be used to express partial similarity between words by overlaps in the binary code. If this binary code is provided in varying length (Owoputi et al. 2013), then the tagging accuracy improves during training to a greater extent than just by providing the entire string as a cluster ID. Brown clustering is a hard-clustering algorithm, and a word will eventually be part of only one cluster. This contrasts with soft-clustering algorithms, such as *Latent Dirichlet Allocation (LDA)* (Blei et al. 2003; Chrupala 2011), which uses probabilistic word classes, and with which a word can belong to more than one cluster. Horsmann and Zesch (2015) show that Brown clustering is more suitable than LDA for PoS tagging of CMC data.

3 POS TAGGING CMC FROM THE PERSPECTIVES OF LANGUAGE TECHNOLOGISTS AND THE LINGUISTS

3.1 The language technologist's view

From a technical viewpoint, a PoS tagger performs well if it reaches a high accuracy and is robust against transfers to other domains of textual data. This high accuracy is a criterion readily fulfilled by many tagger implementations, while the criterion of robustness is often not. Taggers are usually evaluated by choosing one corpus and splitting it up into a training and testing set. The most prominent example of this approach for English is the same corpus evaluation of the Wall Street Journal (WSJ) (Marcus et al. 1993) based on a de-facto standard

data split. Each new tagger implementation reports the tagging results on this data split as point of reference to other implementations. Such evaluations reach high accuracies, but they also evaluate under ideal conditions, since the training and testing data are very similar to each other (Giesbrecht and Evert 2009). This high similarity is unrealistic for real setups, however, and as soon foreign domain data is used for such evaluations the tagging accuracy decreases, with the severity of this decline depending on the degree of dissimilarity. The CMC domain is a such a severe case, with the Stanford tagger (Toutanova et al. 2003), for instance, achieving over 97% accuracy with the WSJ data (Manning 2011), but only 80% with the CMC data set examined by Ritter et al. (2011).

It thus seems as if there is no all-round tagger within reach, as no newswire-trained tagger has a sufficiently high robustness to work on the CMC domain with a similar high accuracy as that seen on edited standard-text. This lack of robustness has motivated considerable research into domain adaptation to re-train tagger models on a mixture of data from several domains, and provide supplementary knowledge from other resources.

3.2 The linguist's view

For qualitative and quantitative empirical analyses of authentic language data, linguists are interested in using corpora which provide highly accurate PoS annotations, and can thus be queried not only for word tokens, but also for morphosyntactic patterns. For the domain of edited text (fictional prose, scientific and newspaper text and similar genres), the reference corpora provided by the Berlin-Brandenburg Academy of Sciences (DWDS corpus, Geyken 2007) and by the Institute for the German Language (DEREKO, Kupietz et al. 2010) are examples which meet this requirement. For the domain of CMC, corpora with highly accurate linguistic annotations still need to be developed, since existing taggers still cannot sufficiently deal with the linguistic peculiarities of CMC discourse.

From a linguistic perspective, and especially for research on the commonalities and differences between the written, interactional language of CMC, the written language of edited text and the language of spoken interactions, a PoS layer in CMC corpora should, on the one hand, adequately represent units which are specific to CMC discourse – such as emoticons, hashtags, non-inflected verb stems (*grins*, *lach*, *grübel*), addressing terms, email addresses and URLs. On the other hand, taggers should also be able to deal with phenomena which are not unique to CMC data but are typical for all types of discourse in informal, interactional settings with spontaneous language production. Besides CMC genres, phenomena of that type occur in spoken language and even in certain domains of edited text (e.g. in direct

speech or quotations as parts of literary prose or newspaper articles). Examples of phenomena of this type are interjections, discourse markers, modal particles and intensifiers, colloquial contractions, and onomatopoeia – phenomena which are only rudimentarily covered by PoS tagsets which have been created for processing edited or newswire texts. The Stuttgart-Tübingen Tagset (STTS, Schiller et al. 1999) for instance, which is a de-facto standard for the tagging of German text corpora, includes a tag for interjections (ITJ), whereas modal particles, downtoners, intensifiers, focus and gradation particles are not represented as unique categories (instead, they are included in the ADV category for adverbs). For contractions, the tagset only covers preposition-article contractions (APPRART) which are part of the written standard, and which are characterised by a high degree of grammaticalisation (German *im, am, zum, vom, ins*); the vast variety of contractions beyond the APPRART type which are typical of colloquial language (e.g., verb-pronoun, conjunction-pronoun, adverb-article) cannot be adequately labelled using STTS.

A precise PoS annotation which covers the aforementioned phenomena can, moreover, form the basis for the (manual or NLP-assisted) creation of more sophisticated corpus annotations, e.g. on syntactic, semantic, pragmatic or interactional patterns.

4 **EMPIRIST: A COMMUNITY SHARED TASK FOR POS TAGGING GERMAN CMC DATA**

In this section, we give a summary of the design and results of a community shared task which was organised to foster the adaptation of NLP tools for the automatic annotation of German CMC data. *EmpiriST* (“Empirikom Shared Task”) resulting from an initiative of the interdisciplinary scientific network “Empirical Research on Computer-mediated Communication” (Empirikom, <http://www.empirikom.net>) which was funded by the DFG 2010–2014, and in which linguists, language technologists, computer scientists and psychologists worked on solutions for open issues related to the acquisition, design and analysis of CMC data sets. A detailed documentation of the task including descriptions of the participating systems is given in WAC-X/EmpiriST (2016).

4.1 **Focus and layout of the task**

The focus of EmpiriST was on PoS tagging of German CMC data in two types of resources: (1) as part of genuine CMC corpora, (2) as part of large corpora




crawled from the web (web corpora). The task provided annotated data sets of CMC and web text to participants as training data to adapt PoS taggers to the CMC domain. EmpiriST consisted of the two subtasks, (1) tokenisation and (2) PoS tagging. These subtasks were performed on two data sets: (i) a CMC data set with samples from several CMC genres (tweets, chats, Wikipedia talk pages, WhatsApp interactions, blog comments), and (ii) a web corpora data set of CC-licensed web pages (including a small portion of CMC discourse). All in all, 23k tokens of training and testing data were annotated, each subset by at least two trained annotators.

4.2 Tagset

EmpiriST adopted the ‘STTS 2.0’ tagset (Beißwenger et al. 2015), which expands the canonical version of the Stuttgart-Tübingen-Tagset (Schiller et al. 1999, henceforth ‘STTS 1.0’) with 18 new tags that are relevant for the tagging of linguistic peculiarities in written CMC interactions that cannot be adequately handled with the STTS 1.0 categories (Table 1). According to the linguist’s view described in Section 3.2, STTS 2.0 introduces two ‘families’ of new tags:

- (i) tags for phenomena that are specific to CMC discourse: ASCII emoticons and emojis, ‘interaction words’ describing facial expressions, gestures, bodily actions, or virtual events (cf. Beißwenger et al. 2012: 3.5.1.3), hashtags, addressing terms, URLs and e-mail addresses.
- (ii) tags for phenomena that are typical of spontaneous (spoken or ‘conceptually oral’) language in colloquial registers: tags for types of colloquial contractions which frequently occur in German chats, tags for discourse markers and onomatopoeia, and, finally, three tags which allow for the description of different types of particles which in STTS 1.0 are treated as adverbs without further subclassification:
 - a tag for intensifiers, focus and gradation particles (which – besides units that belong to the written standard (*sehr, höchst, nur*) – also covers forms which are associated with colloquial registers (*voll geil, krass unterschiedlich*)),
 - a tag for modal particles and downtoners (*Das ist ja / vielleicht doof*),
 - a tag for particles which are part of multi-word lexemes (*keine mehr, noch mal*).

Table 1: Tagset extensions for CMC phenomena according to STTS 2.0.

PoS tag	Category	Examples
I. Tags for phenomena specific for CMC / social media discourse:		
EMO ASC	ASCII emoticon	:-) :-(^^ O.O
EMO IMG	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://www.uni-due.de
EML	E-mail address	peterklein@web.de
II. Tags for phenomena typical for spontaneous (spoken or conceptually oral) language in colloquial registers:		
VV PPER	Tags for types of colloquial contractions which are frequent in CMC (APPRART already exists in STTS 1.0)	schreibste, machste
APPR ART		vorm, überm, fürn
VM PPER		willste, darfst, musste
VA PPER		haste, biste, isses
KOUS PPER		wenns, weils, obse
PPER PPER		ichs, dus, ers
ADV ART	son, sone	
PTK IFG	Intensifier, focus and gradation particles	<u>sehr</u> schön, <u>höchst</u> eigenartig, <u>nur</u> sie, <u>voll</u> geil
PTK MA	Modal particles and downtoners	Das ist ja / <u>vielleicht</u> doof. Ist das <u>denn</u> richtig so? Das war <u>halt</u> echt nicht einfach.
PTK MWL	Particle as part of a multi-word lexeme	keine <u>mehr</u> , <u>noch</u> mal, <u>schon</u> wieder
DM	Discourse markers	<u>weil</u> , <u>obwohl</u> , <u>nur</u> , <u>also</u> , ... with V2 clauses
ONO	Onomatopoeia	boing, miau, zisch

STTS 2.0 is downward compatible to STTS 1.0, and therefore allows for interoperability with existing corpora and tools. In addition, the tagset extensions in STTS 2.0 are compatible with the STTS extensions defined at IDS Mannheim for the PoS annotation of FOLK, the Mannheim “Research and Teaching Corpus of Spoken German” (Westpfahl and Schmidt, 2016). Further details and examples for the tag categories introduced in STTS 2.0 are given in Beißwenger et al. (2015).

4.3 Results for the subtask of PoS tagging the CMC data set

Six teams submitted results for the PoS subtask from eight different systems. The subtask was evaluated in terms of the accuracy of the PoS tag assignments in the participants' submissions. For each system, the submitting team could submit up to three different runs, and only the best was considered in the task results. To put the performance of submissions into perspective, three widely used off-the-shelf tools were additionally evaluated as baselines: TreeTagger v3.2 (Schmid 1995), Stanford tagger v3.6.0 (Toutanova et al. 2003), and the COW pipeline (Schäfer and Bildhauer 2012, Schäfer 2015). Agreement was calculated (1) for the official gold standard on the basis of STTS 2.0, and (2) for the canonical STTS 1.0 on the basis of a coarse-grained mapping of the 18 new tags in STTS 2.0 to the most acceptable corresponding tag(s) in STTS 1.0. The latter was done to allow for a better comparison of the submitted systems with off-the-shelf taggers which are not aware of the STTS 2.0 tagset extensions. Table 2 gives a summary of the results of the submissions and of the three baseline systems for the PoS subtask on the CMC data set. A detailed description of the evaluation metrics and the results is given in Beißwenger et al. (2016).

Table 2: Summary of results of the EmpiriST subtask on PoS tagging for CMC data (Beißwenger et al. 2016).

System	acc (STTS 2.0)	acc (STTS 1.0)
UdS-distributional	87.33	90.28
UdS-retrain	86.40	89.07
UdS-surface	86.45	89.28
LTL-UDE	86.07	88.84
AIPHES	84.22	87.10
bot.zen (<i>non-competitive</i>)	85.42	87.47
\$WAGMOB (<i>non-competitive</i>)	84.77	87.03
COW (<i>baseline</i>)	77.89	81.51
TreeTagger (<i>baseline</i>)	73.21	76.81
Stanford (<i>baseline</i>)	70.60	75.83

The improvements shown by the submitted systems compared to the baseline systems is striking: the best submitted tagger achieved an accuracy of 87.33% evaluated against STTS 2.0 (vs. 77.89% baseline), and an accuracy of 90.28% against STTS 1.0 (vs. 81.51% baseline). Nevertheless, since the EmpiriST training and testing data sets were compiled of snippets of authentic CMC interactions, the number of occurrences of the 18 newly introduced PoS tags in STTS 2.0 was extremely varied, as shown in Table 3.

Table 3: All 18 newly introduced PoS tags from STTS 2.0 with their frequency of occurrence in the training data compared to the frequency of the 18 least frequent STTS 1.0 tags (Horsmann and Zesch 2016).

Tags specific of STTS 2.0	Freq	Least frequent tags in STTS 1.0	Freq
EMOASC	115	PTKANT	42
PTKMA	103	PWAV	39
PTKIFG	99	KOKOM	28
AKW	49	XY	28
HST	46	PDAT	28
ADR	35	VAINF	26
PTKMWL	28	PWS	23
EMOIMG	22	VVIMP	18
URL	18	TRUNC	12
VVPPER	7	KOUI	10
VAPPER	4	PWAT	8
DM	3	VVIZU	7
VMPPER	1	PIDAT	7
ADVART	1	PTKA	5
KOUSPPER	1	APZR	5
ONO	1	VMINF	3
PPERPPER	1	VAPP	3
EML	0	VMPP	1

From the view of corpora representing natural language, the uneven distribution of occurrences with regard to the PoS categories is a notable feature. From the view of language technology, it is an issue that has to be addressed.

4.4 Discussion of the results from the language technologist's perspective: The challenge of rare phenomena

Evaluations of PoS taggers usually focus on the accuracy computed over all PoS tag classes as the main metric of assessment. The frequency of the individual PoS tags varies greatly, which is why a high level of correctness with regard to frequent tags will automatically lead to a high accuracy. At least for English and German, those classes are typically nouns, verbs, adjective and adverbs. Conversely, errors in tagging infrequent tag classes barely have an influence on the accuracy, and thus an accuracy in the mid-nineties tells us little about the system's performance on infrequent tags. More suitable measures do exist, computed for each individual tag, such as the F-score. However, the convenience of having a single value which expresses the overall performance makes accuracy the preferred metric of evaluation.

PoS tagsets for the CMC domain tend to add additional PoS tag classes (Rehbein, 2013, Beißwenger et al. 2015) to address the phenomena of informal language use. Some of these additional tag classes are extremely infrequent, which makes it difficult for the tagger to learn to recognise them during model training. In particular when CMC corpora which ought to represent a certain (sub-)domain are compiled, the problem of infrequency becomes more extreme when tags occur only once or twice. Horsmann and Zesch (2016b) show that such ultra-rare phenomena are not learned by a tagger, even if it is able to reach an accuracy of around 90%.

The lesson learned from the EmpiriST shared task is that annotation of rare phenomena is only reasonable when a sufficient number of samples can be provided for each tag. This certainly conflicts with the goal of having a corpus that represents the natural distribution in a domain. Under practical considerations, when rare phenomena need to be studied, it is more reasonable to give up on the natural distribution and provide additional annotated sequences with the phenomena of interest in order to provide enough training instances to be learned by the tagger.

5 EXPERIMENTS IN POS TAGGING LOW-FREQUENT LINGUISTIC PHENOMENA: THE CASE OF GERMAN VERB-PRONOUN CONTRACTIONS

In this section, we present an experiment in which we investigate how to improve the tagging accuracy on German **verb-pronoun contractions**. Verb-pronoun contractions belong to the class of phenomena which are not unique to CMC discourse, but typical for spontaneous – spoken or ‘conceptually oral’ – language in colloquial registers. Phenomena of this type are of special interest to linguists who want to use corpora to compare written discourse from the CMC domain with the language of edited text and that found in informal, spoken interactions. Table 4 shows examples of such contractions taken from the Dortmund Chat Corpus (Beißwenger 2013, Lungen et al. 2016). Compared to other PoS classes, verb-pronoun contractions must be considered a rarely occurring phenomenon; at the same time, the number of possible forms for this pattern that may occur in a corpus cannot be predicted. In the EmpiriST training data, we found 12 occurrences (seven of the type full verb + pronoun, four of the type auxiliary + pronoun, one of the type modal verb + pronoun, cf. Table 3). Since the use of verb-pronoun contractions is considered typical for informal settings, the frequency of its occurrence may vary in different CMC genres and contexts (e.g., social chats vs. chats in the context of learning and teaching). Verb-pronoun contractions are

therefore an excellent case to explore how a tagger can be adapted to the identification of phenomena which typically (1) occur rarely, (2) in a big variety of possible forms, and without (3) the number of occurrences and the variety of forms being able to be anticipated.

Table 4: Examples of contractions of a full verb with a personal pronoun.

wiederholen (to repeat) + es (it)	1st person
<p>ich wiederhols nochmal, ihr redet hier öffentlich! <i>I repeat it [repeat-it] again, you're talking in public!</i></p>	
kommen (to come) + du (you)	2nd person
<p>wieso? wo kommste denn her? ich besuch dich auch! <i>why? where do you come [come-you] from? i will visit you too!</i></p>	
finden (to find) + du (you)	2nd person
<p>nö,dat ebste findeste eigentlich wenn du gar nich suchst sondern einfach guckst was da ist <i>nope, you find [find-you] the best when you're not searching for it but just look what's there</i></p>	
machen (to make) + es (it)	3rd person
<p>shortnews.de machts möglich wenn die supermarktwebcams reinverlinkt werden:-) <i>shortnews.de makes it [makes-it] possible when they link to the super market webcams:-)</i></p>	

As a prerequisite for studying the use of this phenomenon in the CMC domain, we are adapting a tagger for dealing with VVPPER contractions so that it may be used as a tool for retrieving new instances of VVPPER in raw data. This tagger needs high precision to avoid screening through countless false positive instances, and at the same time we need to be able to find new lexical instances for our studies, which requires a high level of recall. Building such a tagger needs a sufficiently large number of training instances, which poses the biggest challenge to this project, as such data is not readily available. We will thus address two sub-problems: first, how to deal with the lack of training data, and second, how to reach a reasonable trade-off between precision and recall. The focus of our experiments will lie on verb-pronoun contractions of the type *full verb + personal pronoun*, for which STTS 2.0 introduces the tag **VVPPER** with 'VV' representing the full verb (German *Vollverb*) and 'PPER' the personal pronoun (German *Personalpronomen*) component.

5.1 Data set

For building our training data set, we build on the (small) set of 23k manually PoS annotated tokens provided in the context of the EmpiriST project (cf. Section 4) which was annotated using STTS 2.0 (Beißwenger et al. 2015). There are 13 VVPPER instances in the EmpiriST data set, which we split into the training set (seven occurrences, cf. Table 3) and testing set (six occurrences).

Since the VVPPER tag is not included in the canonical STTS, the low representation of the phenomenon in the data cannot be increased using existing corpora which are tagged with STTS 1.0. Therefore, to arrive at meaningful results, we have to increase the number of verb contractions artificially. To do so, we manually select 230 user posts containing this phenomenon from the Dortmund Chat Corpus and machine-tagged these using the Stanford tagger. We manually assign the correct PoS tag from the STTS 2.0 to all VVPPER occurrences, but leave the remaining tags untouched. We have no interest in reaching a new *best-accuracy result*, and thus the performance on other tags is not of primary importance. Of course, ensuring the correctness of the surrounding tags is desirable, but we want to avoid labour intensive, manual annotation as much as possible. We therefore focus on providing verified lexical (context) knowledge of VVPPER and risk wrong surrounding tags as a result of the machine tagging. This enables us to add many additional sequences and inform the tagger more extensively about the phenomenon of interest. Of the 230 instances, we add one half (115) to the test set and one sixth (38) to the training set. The remaining two sixths (77) are the (held back) development set, and will be used in the experiment to increase the number of instances. Hence, our enhanced data set now contains 45 (7+38) VVPPER instances in the training set (seven from the EmpiriST data set and 38 from the additional chat data set) and 121 VVPPER instances in the test set (six EmpiriST, 115 chat). These should be enough training instances for learning the phenomenon, and enough instances for evaluating the tagger, especially with respect to generalisation.

The set of 230 chat posts with PoS annotations can be retrieved from the CLARIN repository at IDS Mannheim via <http://hdl.handle.net/10932/00-0374-4A34-CED0-0801-B> and may be re-used by developers under a CC-BY-SA license.

5.2 PoS Taggers

To find the system which is best suited to the task, we experiment with various PoS taggers and compare different tagger implementations to each other:

Stanford: We include the Stanford (Toutanova et al. 2003) tagger as a widely-used system and train maximum entropy models. We use the default configuration provided for training the German STTS (1.0) model.

HunPos: A Hidden-Markov model based tagger by Halácsy et al. (2007) which is a freely available re-implementation of the TnT tagger by Brants (2000). We choose this tagger to have a further well-known tagger in our setup which is frequently used in the literature, and thus to provide a comparison with the results achieved with the Stanford tagger.

LSTM: A deep-learning PoS tagger by Plank et al. (2016) which is based on Long-Short-Term-Memory (Hochreiter and Schmidhuber 1997) neural networks. This tagger has an interesting property, as it considers the word frequency during model training, which leads to an improved performance on rare words. For our purposes, we argue that rare words and the tagging of rare tags are highly related, as rare tags often also have only rarely occurring word forms. This particular implementation might thus offer some advantages for our use case. We run the tagger with the same parametrisation as Plank et al. (2016), and use a German word embedding which we create from 195 million tokens of German Twitter messages we crawled between 2011 and 2017.

Two-Step: Horsmann and Zesch (2016a) proposed a tagger architecture for CMC data that first uses a highly generalised *coarse-grained* tagger, and as a second step applies a specialised non-sequential tagger for *fine-grained* tagging. The second tagger is tailored towards recognising the tag of interest, while the first tagging step constrains the second tagger, e.g. the non-sequential tagger fitted to verbs contractions would be only applied if the sequence model has tagged a word as a verb. We train the coarse-grained sequence tagging model by using Conditional Random Fields (Lafferty et al. 2001) on the abovementioned training set of EmpiriST data and additionally annotated VVPPER instances. The STTS 2.0 tags are mapped to the coarse-grained tagset by the Universal Dependencies project. We add mappings for the contraction phenomena which are not part of the canonical STTS, and treat the VVPPER instances as a verb form. We include a PoS dictionary and Brown clusters (Brown et al. 1992) created from German Twitter messages to compensate for the lack of CMC training data. This coarse tagger reaches an F1 of 0.93 on the coarse-tag *Verb* in the test data set, which is essential for tagging VVPPER.¹

We train a Support Vector Machine (SVM) for the second step using Weka (Hall et al. 2009), a machine learning toolkit. The SVM is trained on the same data as the sequence model, and is fitted to the local word context in which the VVPPER instances occur. As context features, we use the current word and the first and second words to the right and left. We also use character bigrams over all verbs.

¹ As such, some VVPPER instances might be missed if the coarse-model does not predict 'verb'.

5.3 Experiment: Frequency weight vs. lexical knowledge

In this experiment, we want to learn which information is more relevant for tagging VVPPER instances. We experiment with altering the frequency in the training data by over- and undersampling, and compare the performance to when adding newly annotated instances.

Setup: While annotation of more data will certainly improve the performance, we also want to investigate if we can improve tagging of this particular PoS tag by altering the overall tag distribution. This can either be done by **oversampling** the few instances in the data set (cf. weighting of data, Daumé III, 2007) or by **undersampling**, i.e. removing data from the large other PoS tag classes. Both approaches lead to an increased frequency weight of the focal phenomenon by increasing its frequency relative to the rest of the corpus. If undersampling is applied, sentences which *do not* contain the tag of interest are removed. This shrinks the overall corpus size, so that the tag becomes more frequent than in the original distribution. If oversampling is applied, the sentences with the phenomenon are added several times to increase its frequency weight, but leaving the rest of the corpus untouched. We use the following sampling levels:

- *Downsampling:* We remove 25, 50 and 75 percent of the training data instances which do not contain any VVPPER instances.
- *Oversampling/new instances:* To reach comparable results between oversampling and adding new training instances, we constrain the oversampling to fit the number of held back hand annotated sequences. We thus oversample the additionally added training data two and three times and compare this to adding the same amount of newly annotated data from the held back data.

Results: In Figure 1, we show the results on *out-of-vocabulary* (OOV) instances which did not occur in the training set and, hence, show the performance of the taggers to find new lexical forms. We focus on OOV instances because all taggers perform well in recognising *in-vocabulary* words, with an F1 between 0.96 to 0.99. Neither downsampling nor oversampling helps to achieve a substantial improvement on the tag. Furthermore, downsampling shows that the already small amount of training data becomes a large problem for the LSTM if this is further reduced. The Stanford tagger lags behind the other taggers with both sampling methods. Unsurprisingly, the only effective method is providing new data. With this approach, the LSTM needs considerably more data to improve, while the other taggers improve linearly with each new data set.

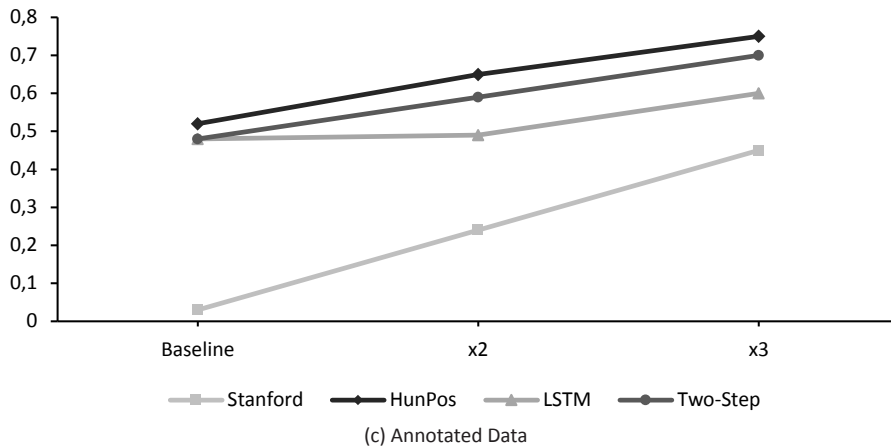
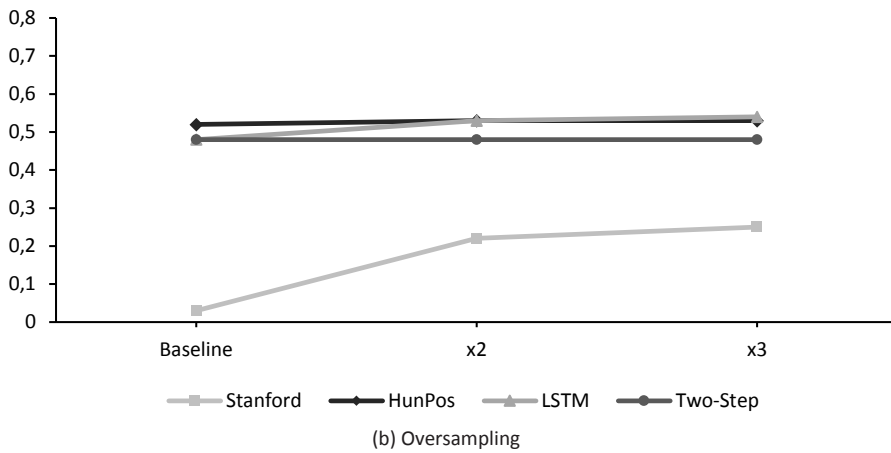
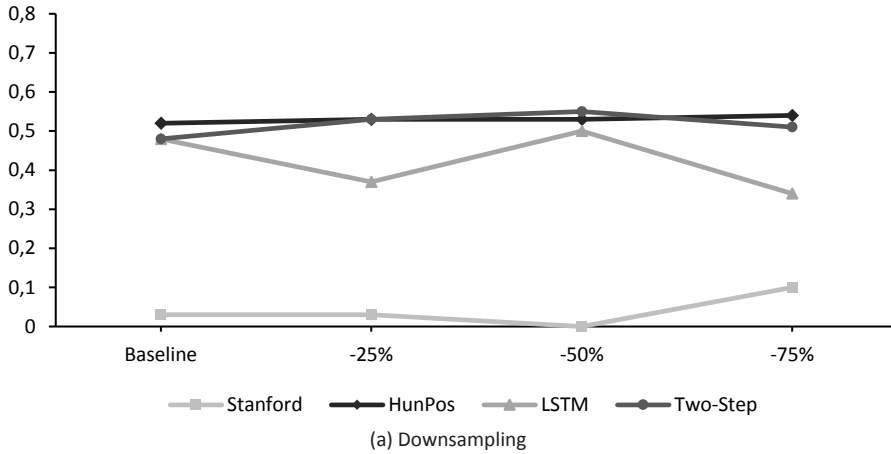


Figure 1: Results on *unknown* VPPER word forms with various methods.

Discussion: Table 5 shows details of the two best taggers, HunPoS and Two-Step. We focus again on the out-of-vocabulary instances, this time presenting also precision (P) and recall (R). The overall F1 score shows that the overall performance of both taggers is rather similar. When looking at precision and recall, highlighted in grey, we see that Two-Step is considerably more precise than HunPos, which has a better recall.

Since oversampling showed barely any effect, we suspect that the added lexical knowledge can account for the improvements we see when adding more data. This would mean that the tagger focuses too much on the lexical forms and does not weight the word context sufficiently.

Table 5: F_1 on all and on out-of-vocabulary instances.

		All	Out-Of-Vocabulary		
Setup		F1	P	R	F1
HunPos	Baseline	.78	.80	.38	.52
	Downs. 75%	.78	.63	.48	.54
	Downs. 50%	.79	.74	.41	.53
	Downs. 25%	.79	.81	.40	.53
	Overs. x2	.79	.78	.40	.53
	Overs. x3	.79	.74	.41	.53
	Annotated x2	.83	.80	.56	.65
	Annotated x3	.88	.81	.70	.75
Two-Step	Baseline	.77	.95	.32	.51
	Downs. 75%	.78	.96	.38	.55
	Downs. 50%	.80	.96	.38	.53
	Downs. 25%	.79	.92	.32	.48
	Overs. x2	.77	.95	.32	.48
	Overs. x3	.77	.95	.32	.48
	Annotated x2	.81	.93	.43	.59
	Annotated x3	.85	.92	.56	.69

5.4 Experiment: Forced generalisation

In this experiment, we examine if we can improve the performance of the Two-Step tagger by forcing it to rely more on the local word context, and thus improve the recall. Since this tagger is self-implemented, we can easily adjust the implementation. We alter the feature space of the SVM and exclude all features which contain the lexical form of the positive instances. The SVM is thus not aware of any lexical forms that can occur with the PoS of interest, and must now rely more strongly on the word context.

Results: In Table 6, we show the changes in performance of the contextualised Two-Step tagger. In parentheses, we show the differences compared to the non-contextualised tagger in Table 5. For both setups, we see an improvement on the overall F1, but the recall especially increases for out-of-vocabulary instances. The overall F1 reached by HunPos (.88) is still better, but the trade-off between precision and recall of Two-Step more efficiently supports the use case of using the tagger as a filtering tool.

Table 6: Results of the contextualised Two-Step.

	All	Out-Of-Vocabulary		
Configuration	F1	P	R	F1
Baseline	.81 (+.04)	.93 (+.02)	.41 (+.09)	.57 (+.09)
Annotated x3	.86 (+.01)	.89 (-.03)	.62 (+.06)	.73 (+.04)

5.5 Experiment: Field trial in CMC

So far, we have only simulated our use case of using a tagger as a filtering tool. Now we turn to a real setting: we tag plain CMC data to find VVPPER instances. Working on unlabelled text means that the ground truth for computing the recall is unknown. We will thus focus on evaluating the precision of the tagging and evaluate how many new instances are found. We choose the Twitter domain for its ease of obtaining data, but also for its linguistic diversity. Some tweets may grammatically and orthographically conform to the written standard while others – more similar to social chat than to edited standard-text – may be noisy and deviant from the orthographic standard, and contain conceptually oral and colloquial language. Tweets of the latter type are the kind of data in which we expect occurrences of VVPPER and other types of colloquial contractions. Twitter thus provides us with a text domain which contains a large amount of naturally occurring noise (which, of course, from the linguist’s view, may be the data which is most interesting for analysing the peculiarities of CMC). Evaluating this domain will provide us with a conservative, lower-bound performance for finding this phenomenon. We use the contextualised Two-Step tagger for its higher precision while still providing reasonably high recall.

Twitter Data: We use tweets that we crawled between 2011 and 2017 from the public Twitter API² endpoint, which allows retrieval of a random subsample of all world-wide posted Twitter messages when this endpoint is accessed. We language-filter those tweets and extract a random sample of 50k German tweets

² <https://dev.twitter.com/streaming/public?lang=en>, last accessed 6th of June, 2017.

(about 1.7 million tokens) between the years 2011 to 2017. All occurrences of addressing terms, hashtags and URLs are replaced by a text constant. The tweets are tokenised by Gimpel et al.'s (2011) ArkTools tokeniser.

Tagger setup: We train the coarse model and the SVM on the full EmpiriST data set including the additionally annotated data. To provide more lexical knowledge and increase the robustness when facing standard language text, we also add 100k tokens of the German newswire Tiger (Brants et al. 2004) corpus to both tagging steps.

Evaluation setup: We evaluate the tagged instances with two annotators. The annotators make four distinctions: *strict*, *relaxed*, *all* and *none*. *Strict* are full verb contractions with personal pronoun (VPPER), the exact phenomenon we intended to tag. *Relaxed* counts all verb contractions with a personal pronoun as correct, which also includes contractions with modal and auxiliary verbs as the first component (VMPPER and VAPPER according to STTS 2.0). *All* counts all phenomena as correct which, from a linguistic perspective, can be considered contractions. This additionally includes, for instance, contractions of conjunctions with personal pronouns, of adverbs with articles, or of two personal pronouns. The remaining cases are not contractions, and thus treated as false positives (= *none*).

We evaluate two setups. The first selects the first 250 of all found instances, which is the basis for the overall evaluation. The second evaluation focuses on out-of-vocabulary instances in which we remove all tagged instances that are known from the training set until we have gathered 250 instances. This set of instances is used to evaluate how frequently new instances are found.

Results: On 50k tweets we find 1,091 instances in total in which one word was tagged as VPPER. The two annotators reach perfect agreement on the subset of the first 250 instances that are evaluated manually. Figure 2a shows the precision of the overall evaluation. The *strict* result shows that the majority of found instances are the targeted VPPER contractions. Including modal and auxiliary verbs in the *relaxed* mode, three quarters of all matches are true positives. When considering *any* type of contractions true positives (in *all*), almost all instances are true positives. We also analysed the type³/token ratio, which is 0.33 for the *strict* evaluation, showing that few instances re-occur with high frequency.

In Figure 2b, we take a closer look at the performance of detecting new contractions, e.g. out-of-vocabulary instances. We focus our discussion on the *strict* results where only VPPER instances count as true positives. The precision is

³ Many word-forms differ by an apostrophe and are, thus, distinct types, e.g. *geht's* vs. *gehts* vs. *geht's* which are counted as three types.

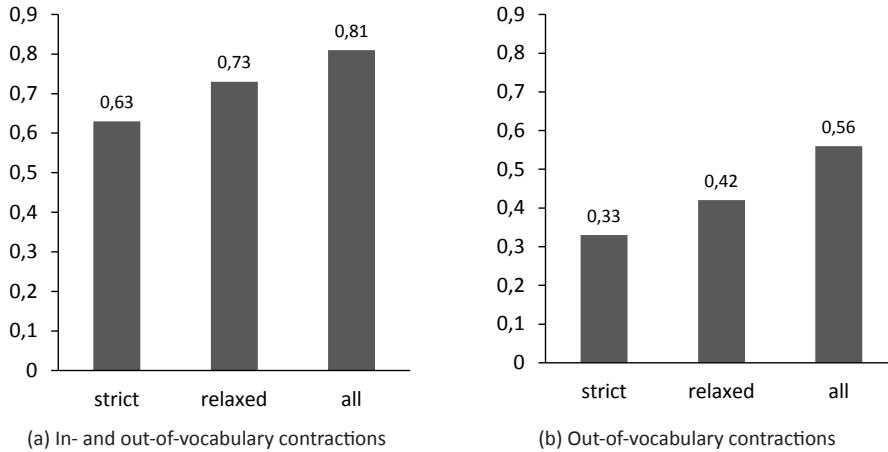


Figure 2: Results of manual evaluation.

drastically reduced to almost half the value when including all instances. The type/token ratio of 0.69 is almost twice as high as the overall evaluation. This confirms that the tagger is able to recognise many new instances of the phenomenon. Furthermore, when ignoring the known instances, almost every correct instance is a new lexical form.

Table 7: Examples of tagged instances (bold) in context and PoS category according to STTS 2.0.

<i>Strict</i>		
	Savegames - jetzt langts aber !	VVPPER
	Da lernste pragmatisch zu sein .	VVPPER
	Ich sachs dir noch .	VVPPER
<i>Relaxed</i>		
	Ich bins auf jeden Fall nicht .	VAPPER
	Wer hats gedacht .	VAPPER
	Ich wills nicht ich will aber auch nicht [...]	VMPPER
<i>All</i>		
	So schlimm hab ich's mir mit noch keiner Ex verscherzt .	PPERPPER
	Warum einfach , wenn's auch kompliziert geht ? URL	KOUSPPER
	Ich beschränke mich auf's nicht im Weg stehen .	APPRART
<i>Frequent Confusion Cases</i>		
	Und keiner weiss warum .	VV
	Ich weiss gar nicht , was du beruflich machst .	VV
	Ich weis wie immer nicht ... URL	VV

Discussion: Table 7 shows examples of each of the three evaluation modes (*strict*, *relaxed*, *all*) and additionally presents three instances of a frequent confusion case which is erroneously tagged as contraction. In the *strict* case there are instances in quite different local word contexts, which supports our motivation for studying this phenomenon. A general observation about the SVM is that it seems to be biased on word endings on <ss> or <’s>. Such words have a high chance of being tagged as contractions. This bias also seems to account for a rather common confusion case with the verb *weiß* (to know), where the German <ß> is erroneously replaced by <ss> but at the same time accounts for the related phenomena in the *relaxed* and *all* evaluation. We are planning to address the further reduction of false positives in future work.

6 CONCLUSION

In view of the heterogeneous frequency of CMC phenomena in CMC data, the results and lessons learned from the EmpiriST shared task suggest that it is not realistic to train a tagger which performs well on any phenomena on the token/PoS level.

In particular, finding rare or ultra-rare phenomena poses serious challenges, and the small size of hand-annotated CMC training data sets causes the under-representation of such phenomena. The EmpiriST project conducted by Beißwenger et al. (2016) showed that the degree of under-representation can be so severe that machine learning methods fail almost entirely to learn how to recognise these phenomena. Increasing the frequency of rare phenomena artificially by over- and undersampling has no impact on this, as the phenomena occur just too infrequently. We thus presented a case study in which we used a PoS tagger as a filtering tool to find instances of German verb-pronoun contractions. We started from the EmpiriST training data and added an additional set of 230 hand-annotated user posts which had been selected manually from the Dortmund Chat Corpus as further instances of the phenomenon of interest. The results shows that the choice of the tagger together with the expansion of the training data with relatively small amounts of additional instances turns out to be a promising way to let the tagger learn the local word context, and thus enables tagging such phenomena with a sufficiently high recall and precision. To reduce the number of false positives, we are planning to add the results of the manual evaluation of the first 250 positives found in tweets to our training data set, and then retrain the SVM on the expanded data in a bootstrapping approach. In future work we will also investigate how tagging improves if not just the instances of interest are hand-annotated, but also their local word context, in order to find the ideal trade-off

between avoiding annotation of full sentences and yet achieving improved results for a certain phenomenon.

To be able to estimate if the results of our case study may provide a general and more efficient approach to “nasty” phenomena in CMC corpora, the study should be repeated for other CMC phenomena which are either rare and/or difficult to handle with approaches from the literature. More close cooperation between language technologists and linguists is thus recommended, as this would enable the creation and annotation of the high-quality samples from CMC corpora which are needed for training.

References

- Bartz, Thomas, 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.
- Beißwenger, Michael, 2013: Das Dortmunder Chat-Korpus. *Zeitschrift für germanistische Linguistik* 41/1. 161–164.
- Beißwenger, Michael, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer and Angelika Storrer, 2012: A TEI Schema for the Representation of Computer-mediated Communication. *Journal of the Text Encoding Initiative* (jTEI) 3. <http://jtei.revues.org/476>. (Last accessed 5 May 2017.)
- Beißwenger, Michael, Thomas Bartz, Angelika Storrer and Swantje Westpfahl, 2015: *Tagset und Richtlinie für das Part-of-Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation / Tagset and guidelines for the PoS tagging of language data from genres of computer-mediated communication / social media. Empirist guideline document (German and English version)*. <https://sites.google.com/site/empirist2015/home/annotation-guidelines>. (Last accessed 5 May 2017.)
- Beißwenger, Michael, Sabine Bartsch, Stefan Evert and Kay-Michael Würzner, 2016: Empirist 2015: A Shared Task on the Automatic Linguistic Annotation of Computer-Mediated Communication and Web Corpora. *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the Empirist Shared Task*. Stroudsburg: Association for Computational Linguistics. 44–56. <http://aclweb.org/anthology/W/W16/W16-2606.pdf>. (Last accessed 5 May 2017.)
- Blei, David M., Andrew Y. Ng and Michael I. Jordan, 2003: Latent Dirichlet Allocation. *The Journal of Machine Learning Research* 3. 993–1022.

- Brants, Thorsten, 2000: TnT: A Statistical Part-of-speech Tagger. *Proceedings of the Sixth Conference on Applied Natural Language Processing*. Seattle: Association for Computational Linguistics. 224–231.
- Brants, Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith and Hans Uszkoreit, 2004: TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation* 2/4. 597–620.
- Brown, Peter F., DeSouza, Peter V., Mercer, Robert L., Pietra, Vincent J. Della, Lai, Jenifer C., 1992: Class-Based n-gram Models of Natural Language. *Computational Linguistics* 18. 467–479
- Chrupala, Gzegorz, 2011: Efficient induction of probabilistic word classes with LDA. *Proceedings of the Fifth International Joint Conference on Natural Language Processing*. Chiang Mai: Asian Federation of Natural Language Processing. 363–372.
- Chrupala, Gzegorz, 2014: Normalizing tweets with edit scripts and recurrent neural embeddings. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore: Association for Computational Linguistics. 680–686.
- Cook, Paul, Stefan Evert, Roland Schäfer and Egon Stemle (eds.) *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*. Stroudsburg: Association for Computational Linguistics (ACL Anthology W16-26). <http://aclweb.org/anthology/W/W16/W16-26.pdf>. (Last accessed 5 May 2017.)
- Daumé III, Hal, 2007: Frustratingly Easy Domain Adaptation. *Conference of the Association for Computational Linguistics (ACL)*. Czech Republic: Association for Computational Linguistics. 256–263.
- Eisenstein, Jacob, 2013: What to do about bad language on the internet. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta: Association for Computational Linguistics. 359–369.
- Geyken, Alexander, 2007: The DWDS corpus: A reference corpus for the German language of the 20th century. Fellbaum, Christiane (ed.): *Idioms and Collocations: Corpus-based Linguistic and Lexicographic Studies*. London: Bloomsbury Publishing. 23–41.
- Giesbrecht, Eugenie and Stefan Evert, 2009: Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. *Proceedings of the Web as Corpus Workshop (WAC)*. San Sebastian.
- Gimpel, Kevin, 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. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers 2*. Stroudsburg: Association for Computational Linguistics. 42–47.

- Halácsy, Péter, András Kornai and Csaba Oravecz, 2007: HunPos: An open source trigram tagger. *Proceedings of the 45th Annual Meeting of the ACL*. Association for Computational Linguistics. 209–212.
- Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H. Witten, 2009: The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter* 11/1, <http://dl.acm.org/citation.cfm?id=1656278>. (Last accessed 5 May 2017.)
- Han, Bo and Timothy Baldwin, 2011: Lexical Normalisation of Short Text Messages: Makn Sens a #Twitter. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies 1*. Stroudsburg: Association for Computational Linguistics. 368–378.
- Hochreiter, Sepp and Jürgen Schmidhuber, 1997: *Long Short-Term Memory*. *Neural Computation*. MIT Press. 1735–1780.
- Horbach, Andrea, Diana Steffen, Stefan Thater and Manfred Pinkal, 2014: Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication. *Proceedings of KONVENS 2014*. Hildesheim. 171–177.
- Horsmann, Tobias and Torsten Zesch, 2015: Effectiveness of Domain Adaptation Approaches for Social Media PoS Tagging. *Proceeding of the Second Italian Conference on Computational Linguistics*. Trento: Accademia University Press. 166–170.
- Horsmann, Tobias and Torsten Zesch, 2016a: Assigning Fine-grained PoS Tags based on High-precision Coarse-grained Tagging. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka: Dublin City University and Association for Computational Linguistics.
- Horsmann, Tobias and Torsten Zesch, 2016b: LTL-UDE @ EmpiriST 2015: Tokenization and PoS Tagging of Social Media Text. *Proceedings of the 10th Web as Corpus Workshop (WAC-X)*. Berlin: Association for Computational Linguistics. 120–126.
- Kupietz, Marc, Cyril Belica, Holger Keibel and Andreas Witt, 2010: The German Reference Corpus DeReKo: A primordial sample for linguistic research. Calzolari, Nicoletta et al. (eds.): *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010)*. Valletta: European Language Resources Association (ELRA). 1848–1854. http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf. (Last accessed 5 May 2017.)
- Lafferty, John D., Andrew McCallum and Fernando C. N. Pereira, 2001: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco: Morgan Kaufmann Publishers Inc. 282–289.
- Lüngen, Harald, Michael Beißwenger, Axel Herold and Angelika Storrer, 2016: Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. Dipper, Stefanie, Friedrich Neubarth and Heike Zinsmeister (Eds.): *Proceedings of the 13th Con-*

- ference on Natural Language Processing* (KONVENS 2016). 156–164. https://www.linguistics.rub.de/konvens16/pub/20_konvensproc.pdf. (Last accessed 5 May 2017.)
- Manning, Christopher D., 2011: Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing 1*. Tokyo: Springer-Verlag Berlin, Heidelberg. 171–189.
- Marcus, Mitchell P., Mary Ann Marcinkiewicz and Beatrice Santorini, 1993: Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19/2. Cambridge: MIT Press. 313–330.
- Neunerdt, Melanie, Michael Reyer and Rudolf Mathar, 2013: A POS Tagger for Social Media Texts trained on Web Comments. *Polibits* 48. 61–68.
- Neunerdt, Melanie, Michael Reyer and Rudolf Mathar, 2014: Efficient Training Data Enrichment and Unknown Token Handling for POS Tagging of Non-standardized Texts. *12th Conference on Natural Language Processing (KONVENS)*. Hildesheim. 186–192.
- Owoputi, Olutobi, Chris Dyer, Kevin Gimpel, Nathan Schneider and Noah A. Smith, 2013: Improved part-of-speech tagging for online conversational text with word clusters. *Proceedings of the Conference of North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Plank, Barbara, Anders Søgaard and Yoav Goldberg, 2016: Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL-16)*. Berlin: Association for Computational Linguistics. 412–418.
- Rehbein, Ines, 2013: Fine-Grained POS Tagging of German Tweets. *Language Processing and Knowledge in the Web*. Springer-Verlag Berlin, Heidelberg. 162–175.
- Ritter, Alan, Sam Clark, Mausam Etzioni and Oren Etzioni, 2011: Named Entity Recognition in Tweets: An Experimental Study. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Edinburgh: Association for Computational Linguistics. 1524–1534.
- Schäfer, Roland, 2015: Processing and querying large web corpora with the COW14 architecture. Bánski, Piotr, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lungen and Andreas Witt (eds.): *Proceedings of Challenges in the Management of Large Corpora 3*. Lancaster: UCREL.
- Schäfer, Roland and Felix Bildhauer, 2012: Building large corpora from the web using a new efficient tool chain. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC '12)*. Istanbul: ELRA. 486–493.

- Schiller, Anne, Simone Teufel, Christine Stöckert and Christine Thielen, 1999: *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. Stuttgart: Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung. <http://www.sfs.unituebingen.de/resources/stts-1999.pdf>. (Last accessed 5 May 2017.)
- Schmid, Helmut, 1995. Improvements in part-of speech tagging with an application to German. *Proceedings of the ACL SIGDAT Workshop*.
- Toutanova, Kristina, Dan Klein, Christopher D. Manning and Yoram Singer, 2003: Feature rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Universal Dependencies 1.2*. Universal Dependencies Consortium. <http://universaldependencies.github.io/docs/>. (Last accessed 5 May 2017.)
- van Halteren, Hans and Nelleke Oostdijk, 2014: Variability in Dutch Tweets. An estimate of the proportion of deviant word tokens text. *Journal of Language Technology and Computational Linguistics (JLCL)* 29/2. 97–123.
- Westpfahl, Swantje and Thomas Schmidt, 2016: FOLK-Gold – A GOLD standard for Part-of-Speech- Tagging of Spoken German. *Proceedings of the Tenth conference on International Language Resources and Evaluation (LREC16)*. Paris. 1493–1499.