

# Discussing best practices for the annotation of Twitter microtext

Ines Rehbein, Emiel Visser  
and Nadine Lestmann

E-mail: irehbein|visser|lestmann@uni-potsdam.de

## Abstract

This paper contributes to the discussion on best practices for the syntactic analysis of non-canonical language, focusing on Twitter microtext. We present an annotation experiment where we test an existing POS tagset, the Stuttgart-Tübingen Tagset (STTS), with respect to its applicability for annotating new text from the social media, in particular from Twitter microblogs. We discuss different tagset extensions proposed in the literature and test our extended tagset on a set of 506 tweets (7.418 tokens) where we achieve an inter-annotator agreement for two human annotators in the range of 92.7 to 94.4 ( $\kappa$ ). Our error analysis shows that especially the annotation of Twitter-specific phenomena such as hashtags and at-mentions causes disagreements between the human annotators. Following up on this, we provide a discussion of the different uses of the @- and #-marker in Twitter and argue against analysing both on the POS level by means of an at-mention or hashtag label. Instead, we sketch a syntactic analysis which describes these phenomena by means of syntactic categories and grammatical functions.

## 1 Introduction

Through the emergence of new technologies, human communication practices have undergone radical changes (examples are communication by email, chat, text messages or Twitter microblogs). New text types from the web, in particular from the social media, challenge traditional views on the distinction between orality and literacy [9, 8] by combining features of both, oral and written communication. Our interest is in understanding these changes and in investigating the properties of these newly emerging text types. For this undertaking, linguistically annotated corpora would be of great help.

However, most annotation schemes for annotating part-of-speech (POS) tags and syntax have been developed for canonical written text (often from the newspaper domain), and it is not clear whether they also allow us to adequately describe the properties of user-generated content from the web.

Furthermore, recent work on POS tagging Twitter data has shown a low agreement of human annotators on tweets, yielding inter-annotator agreement (IAA)

scores in the range of 92-93% while the same scores on canonical, written text are in the high nineties [5]. This is highly problematic for the development of automatic, supervised methods for POS annotation of CMC, as those rely on the quality of the manually annotated data for training and evaluation, which will provide an upper bound on the performance of automatic methods.<sup>1</sup> Thus, to improve the quality of automatic POS tagging of CMC, we need linguistically sound annotation schemes which can be applied with high reliability by the human coders, and which provide a meaningful analysis of the phenomena of CMC.

In the paper, we present an annotation experiment where we assign parts-of-speech from the Stuttgart-Tübingen Tag Set (STTS) [15] to German microtext from Twitter, asking the following questions:

- What are the main problems for analysing computer-mediated communication (CMC) on the POS level, using an annotation scheme developed for canonical written language?
- How reliable are human annotations of POS on social media text?

We report on inter-annotator agreement results obtained for POS tagging German tweets and discuss the reasons for the lower agreement obtained on Twitter microtext as compared to, e.g., newspaper text. Based on our annotation study, we address the main problems encountered during the annotation and propose a different approach, which, in our opinion, is more promising to yield a reliable and adequate analysis of social media data. In particular, we focus on Twitter-specific phenomena like the @- and #-marker.<sup>2</sup> We illustrate that both have multiple functions and argue against an analysis of references and linking information by means of an at-mention and hashtag label on the POS level, as proposed in the literature [14, 6, 3]. Instead, we advocate for encoding this type of information on the syntactic level and sketch a possible solution.

The paper is structured as follows. In section 2, we review related work on extending POS tagsets for the annotation of Twitter microtext. Section 3 presents our annotation experiment and reports on our inter-annotator agreement for human annotation of POS on German tweets. In section 4 we illustrate the problems we encountered during the annotation and discuss different solutions. We conclude in section 5.

## 2 Related work

There is some recent work on developing or extending POS tagsets for annotating Twitter microtext. Ritter et al. [14] expand the Penn treebank POS tagset by adding

---

<sup>1</sup>The relatively low IAA scores for POS-tagging microtext also put into question results for semi-supervised and unsupervised POS tagging as those are evaluated against a (less accurate) hand-crafted goldstandard.

<sup>2</sup>The at-mentions (@) identify the addressee of the tweet and provide a link to the users' Twitter profile while the hashtags (#) function as semantic tags or keywords.

four new, unambiguous tags for hyperlinks, user names, hashtags and retweets and manually annotate a testset of 800 English tweets.<sup>3</sup> They do not give numbers for inter-annotator agreement of the annotations.

Gimpel et al. [6] annotate English tweets using a coarse-grained tagset (25 tags) with five new tags for CMC-specific phenomena. These include emoticons, hyperlinks, hashtags (linking the tweet to a semantic category), at-mentions (indicating the recipient of the tweet) and a tag for annotating 'RT' and ':' in retweet constructions. In contrast to [14], Gimpel et al. [6] annotate tokens as hashtags only when they are not integrated in the tweet message. Syntactically integrated instances of hashtags are annotated according to their distribution. They report an inter-annotator agreement of 0.914 on a small testset of 72 tweets.

Avontuur et al. [1] annotate Dutch tweets, using a hierarchical morpho-syntactic tagset with 320 tags, based on a tagset developed for written text, and the five new, Twitter-specific tags from [6]. They obtain an inter-annotator agreement in the range of 0.912 to 0.933 (Cohen's  $\kappa$ ) on a testset of 1,056 tweets.

In all three studies, the agreement for human annotations on Twitter is in the same range, and substantially lower than the one obtained on canonical, written text (see, e.g., Brants [5] for IAA on German newspaper text).

### 3 Reliability of manual annotations on German tweets

In our annotation experiment, we use the 54 tags of the Stuttgart-Tübingen Tag Set (STTS) [15] to annotate German tweets. We follow the proposals above and also introduce new tags for annotating emoticons, hashtags, at-mentions and hyperlinks. Similar to [6], we only annotate tokens as hashtags when they are not integrated in the tweet. In contrast to [6], we do the same with at-mentions and hyperlinks (also see section 4). As we are interested in investigating conceptual orality in a written register [8], we also add new tags for discourse phenomena such as filled pauses, question tags or backchannel signals from an extension of the STTS developed for annotating spoken language [12] (for details see section 4.1).

A prominent feature of CMC taken from spoken language is the contraction of individual lexical units into a new form, inspired by their pronunciation in spoken discourse. We do not correct these non-canonical tokenisations but follow the approach of Gimpel et al. [6] and use combinations of POS tags to annotate the contracted word forms, as shown in Example (1).<sup>4</sup>

In the experiment, we annotated German tweets which we collected from Twitter over a time period from July 2012 to February 2013, using the Python Tweepy module<sup>5</sup> as an interface to the Twitter Search API<sup>6</sup>. Our test set includes 506 tweets

<sup>3</sup>By "unambiguous" we mean that all tokens starting with an @ or # as well as all hyperlinks and emoticons are labelled with the corresponding tag, regardless of their distribution.

<sup>4</sup>The same approach is taken in the STTS for annotating merged prepositions (APPR) and definite determiners (ART) as APPRART, e.g. *in/APPR dem/ART* (in the) vs. *im/APPRART* (in\_the).

<sup>5</sup><http://pythonhosted.org/tweepy/html>

<sup>6</sup><https://dev.twitter.com/docs/api/1/get/search>

(7,425 tokens) which were annotated independently by two human annotators. Table 1 shows our inter-annotator agreement on the data.

	# Tagset	# Testset	$\kappa$
<i>this work</i>	72	506 tweets	0.92.6 - 0.94.4
<i>Gimpel et al. (2011)</i>	25	72 tweets	0.914
<i>Avontuur et al. (2012)</i>	325	1056 tweets	0.912 - 0.933

Table 1: Inter-annotator agreement on German, English and Dutch tweets.

Our results are in line with other studies on inter-annotator agreement of English and Dutch Twitter data [6, 1]. Most interestingly, the size of the tagset does not seem to have a huge impact on the results. All three studies show an agreement well above 0.9 ( $\kappa$ ), despite the different sizes of the three tagsets.

We now come to the question why IAA on Twitter microtext is so much lower than the one on canonical written text. Our error analysis shows that the most difficult decisions during the annotation concern the distinction between proper names (NE) and nouns (12% of all disagreements), NE and foreign language material (6.3%), NE and at-mentions (5.1%) and NE and hashtags (3.0%).

We take this as a starting point to have a closer look at the disagreements on CMC-specific phenomena and discuss these in more detail. We argue that the POS tags for at-mentions and hashtags do not provide an adequate description of the different functions of these markers and that their linking function should not be encoded on the POS level.

## 4 Twitter – the data

Communication on Twitter is shaped by a liberal use of orthographic rules where spelling conventions are often ignored and the capitalisation of German nouns is not done in a systematic way (1). In addition, German compound words are often split up into their components while, at the same time, individual lexical units are contracted into a new form, inspired by their pronunciation in spoken language.

- (1) der **briten regierung hamse** doch ins gehirn geschissen und vergessen umzurühren  
 the British governm. have\_they but in\_the brain shat and forgotten to stir  
 “The British government got shit for brains”

### 4.1 Features from spoken language

Besides contractions, we find many other features imitating informal spoken language in a written medium. We annotate those using an extended version of the STTS developed for the annotation of spoken language phenomena [12].

One phenomenon is the use of disfluencies like repairs and filled pauses in Twitter which, considering that the communication is not subject to time-pressure caused by online processing and that the users have the possibility to revise and

edit their messages, is at least unexpected.<sup>7</sup> We assign filled pauses in Twitter the PTKFILL tag (2).

- (2) On the road **äh**<sub>PTKFILL</sub> train **äh**<sub>PTKFILL</sub> also Ihr wisst schon :)  
 On the road uh train uh well you know already :)  
 "On the road uh train uh, well, you know :)"

Private communication on Twitter is highly informal, which is shown by the high number of interjections, discourse markers and verbless sentences. Tweets are also highly interactive, as indicated by the frequent use of backchannel signals and question tags which we assign the labels PTKREZ (3) and PTKQU (4).

- (3) @userA: yaa dann mach das soo @userB: **hmm**<sub>PTKREZ</sub> .... muss noch nachdenken !  
 @userA: yaa then do it like that @userB: hmm .... have to still think !  
 " @userA: Yeah, then do it like this @userB: hmm ... still have to think!"
- (4) geil , **wa**<sub>PTKQU</sub> !? xD  
 cool , what !? xD  
 "Cool, isn't it?"

Other extensions cover the use of discourse-structuring particles, onomatopoeia and forms of echoism, unfinished words, and a new punctuation sign for marking abandoned utterances. These extensions to the STTS for annotating spoken language differ from the original STTS by way of being defined by discourse-pragmatic criteria instead of morpho-syntactic ones. It could be argued that these distinctions are hard to operationalise and thus should not be encoded on the POS level. For many NLP applications such as Information Retrieval or Named Entity Recognition, discourse particles do not seem to be relevant. For the comparative study of orality in spoken and written discourse, however, these particles can augment the corpus with useful information.<sup>8</sup>

## 4.2 CMC-specific features

### 4.2.1 Emoticons

A major drawback of the written medium is the lack of important channels of non-verbal communication such as mimics, prosody and stress. To make up for this, Twitter users adopt different techniques to express themselves. In addition to a frequent use of interjections and exclamative constructions, we observe the duplication of characters (5) to express emphasis, and the use of uppercased words to indicate shouting (6). Emoticons are another way of expressing emotion in CMC. We follow [14, 6] and introduce a new tag for the annotation of emoticons (EMO) (5),(6).

- (5) **Awww** wie **süüüß** \*o\*<sub>EMO</sub> (6) **Peinlich** , aber **JA ! :-)**<sub>EMO</sub>  
 Aw how sweet \*o\* embarrassing , but yes ! :-)

<sup>7</sup>See [13] for an analysis of the different functions of filled pauses in Twitter.

<sup>8</sup>Another area where this type of information might be useful is Sentiment Analysis/Opinion Mining.

## 4.2.2 Hyperlinks

Hyperlinks in tweets can be positioned either at the beginning or at the end of the tweet, linking the tweet to additional, external information (7), or can be syntactically integrated in the tweet (8). While [14, 6] use a new, unambiguous tag to annotate all hyperlinks, regardless of their distribution, we distinguish between external links (annotated as URL) and syntactically integrated ones (annotated as proper names).

- (7) bei dem Wetter... <http://t.co/ywjSHuhK>  
with the weather... <http://t.co/ywjSHuhk>  
"In weather like this..."
- (8) Hast du eventuell mal mit <http://t.co/EsNtqGku> verglichen ?  
Have you maybe PTCL with <http://t.co/EsNtqGku> compared ?  
"Have you compared it with <http://t.co/EsNtqGku>?"

The annotators' agreement on the distinction between integrated and non-integrated instances in our annotation study was quite low. This is mostly due to the non-systematic use of punctuation and capitalisation in Twitter which makes sentence segmentation difficult. While examples (7) and (8) are straightforward, in examples (9) and (10) it is less clear whether the hyperlinks are integrated or not.

- (9) Neue monatliche Umfrage jetzt online auf unserer Homepage <http://t.co/cvwiJTLA> .  
New monthly poll now online on our homepage <http://t.co/cvwiJTLA> .
- (10) Und ich dachte schon, #Siri hätte mich nicht mehr lieb: <http://t.co/Xclx> -Siri ist toll  
And I thought already, #Siri was REFL not still fond: <http://t.co/Xclx> -Siri is great  
"And I already thought that #Siri wasn't fond of me any more: <http://t.co...> -Siri is great"

Given that hyperlinks are identifiers referring to objects in the world, we argue that it is appropriate to annotate all hyperlinks as proper names on the POS level. While we acknowledge that the linking information might be useful for some applications, we do not think that they justify the introduction of a new part of speech category but would rather shift this type of information to a different level, e.g. including it as a new Named Entity type and encoding it as part of the syntactic annotation, similar to the approach in the TüBa-D/Z [7] (release 8).

## 4.2.3 At-mentions

Originally, the @-sign has been used as an address marker to refer to the addressee of a tweet (or a chat message) (11), but is now also used in a number of other contexts and with different functions.

- (11) @Schebacca ok warum ist das wichtig ???  
@Schebacca ok why is that important ???  
"@Schebacca Ok, why is that important?"

In (12), the @ occurs in isolation, separated by whitespaces, and is used as a local preposition.

- (12) Rest des Tages dann Home-Office , vielleicht im Garten ? (@ Bahnhof Ansbach)  
 Rest of the day then home office , maybe in the garden ? (@ Bahnhof Ansbach)  
 "Home office for the rest of the day, maybe in the garden? (at Ansbach train station)

In contrast, the @ in (13) is not a token of its own but is contracted with a location name, Bad Hersfeld. There are two possible analyses here. First, we could assume that the @ again functions as a preposition and should be separated from the location name by the tokeniser. The second analysis opposes the first one by assuming that *Bad Hersfeld* is a post-modifying NP, and that the sole function of the @-sign is to provide a link to the profile of *Bad Hersfeld* (without having the explicit semantics of a local preposition).

- (13) Danke an die ehem. Medusa Bar @Bad Hersfeld, top Leute und super Stimmung !  
 Thanks to the former Medusa Bar @Bad Hersfeld, great people and super atmosphere !

The second analysis is backed up by cases like (14), where the @ was merged with a proper name but does not license the reading as a preposition. The attempt to replace the @ with a preposition would even result in an ungrammatical utterance.

- (14) ich folge ja nun der @GrinseDame ..  
 I follow PTCL now the @GrinseDame ..  
 "I now follow the @GrinningLady .."

Examples (15) and (16) support our analysis by showing that the users do not conceptualise the @ as a preposition, but combine user names marked by @ with additional prepositions, which - if the first analysis for (13) was correct - should be redundant.

- (15) Warum wird der scheiss tweet **an** @mondmiri nicht gesendet ???  
 Why is the shitty tweet to @mondmiri not sent ???  
 "Why hasn't the shitty tweet to @mondmiri been sent ?"
- (16) Wenn ich **bei** @lidl eine Stunde am Pfandautomaten warten muss geht es  
 When I at @lidl one hour at the deposit redemption machine wait must goes it  
 immer noch schneller als **im** @kaufland  
 always still quicker as in the @kaufland  
 "Even if I have to wait at @lidl for one hour in the queue for the deposit redemption machine, it'll still be faster than at @kaufland"

In conclusion, we argue that the @ is used as an address marker or preposition only in some cases but has lost its original meaning in many others. We thus refrain from separating the @ from the following token and annotating it as a preposition or an address marker. Our main reason for being rather conservative about changing the tokenisation is that separating all @-signs from user or location names would result in a substantial increase in token numbers for CMC corpora, thus leading to an artificially higher type-token ratio (TTR) for CMC as compared to other types of text. This would lead to skewed results for comparative corpus studies of register variation using corpus-linguistic measures like the TTR, sentence length or measures of syntactic complexity (which are often based

on sentence length).<sup>9</sup>

#### 4.2.4 Hashtags

Similar to hyperlinks and at-mentions, hashtags can be syntactically integrated in the tweet message (17), or can be positioned at the beginning or at the end of the tweet, as in (18). Hashtags can be used as keywords or semantic tags to categorise the tweet and thus allow users to search for other tweets of the same category.

(17) Jetzt **#Stromanbieter** **#vergleichen**  
Now suppliers of electric energy compare  
"Compare suppliers of electric energy now"

(18) "Spül [spiel] mir das Lied vom Tod" **#Spülwitze**  
"Wash [play] me the song of death" #washing jokes

The function of hashtags, however, cannot be reduced to semantic tagging. They are frequently used to add an evaluation to the (otherwise neutral) tweet, as in (19).

(19) Laut meiner **#wetterapp** hat es 7 grad **#toocold**  
As per my weather app has it 7 degrees #toocold  
"According to my weather app we have 7 degrees #toocold"

They can also add relevant context information needed for understanding the message of a (highly underspecified) tweet, as in (20).

(20) Hey drückt @ich\_seh\_weiss um 12 die daumen :-)  
Hey press @ich\_seh\_weiss at 12 the thumbs :-)  
"Hey, fingers crossed for @I\_see\_white at 12 :-)"

Some tweets include nothing but a hashtag (21). These often serve as a statement about the general (emotional) state of mind of the Twitter user, in a highly compressed format.

(21) **#übermüdeteresistenzfrühmeckertweet**  
#overtired-it-is-too-early-rant-tweet

Twitter users are also highly register-aware. Sometimes hashtags are simply used because of this, as stated in the self-ironic tweet in (22).

(22) da fehlt noch #tweet #hashtag  
there lacks still #tweet #hashtag  
**#wortedieichsowiesoschongeschriebenhabeimzweifelnochmalaufenglischalshashtaghinterher**  
#words-which-I-anyway-already-written-have-in-doubt-again-on-English-as-hashtag-afterwards  
"The #tweet #hashtag is still missing here. #words-which-I've-already-written-anyway-when-in-doubt-then-I'll-add-them-in-English-to-the-end-of-the-tweet"

---

<sup>9</sup>For the English OCT27 data set [10], separating the @ would result in a seemingly higher token number of 27,896 as compared to 26,594 tokens in the original data set, and an additional segmentation of the # would further increase the number of tokens to 28,316.

Some hashtags include complex inflective constructions (23).<sup>10</sup>

- (23) #mitfreu #superfreu #keinnekrophilenwitzemach  
 #with-you-rejoice<sub>noninflected</sub> #super-rejoice<sub>noninfl</sub> #no-necrophile-jokes-make<sub>noninfl</sub>

Inflective constructions in CMC are often enclosed by asterisks or inequality signs, but users also encode nouns, verb phrases or whole sentences in that way (24). In most cases, the so-marked constructions are not syntactically integrated in the tweet but function as meta-comments, adding information on the emotional state of the user and her environment, or set the stage (25), similar to stage directions in a screenplay.

- (24) \*Vorfreude\* / \*kaffeetasseheb\* / \*hat schokolade gefunden\*  
 \*anticipation\* / \*coffee\_cup\_lift<sub>noninflected</sub>\* / \*has chocolate found\*
- (25) \*Trommelwirbel\* / \*an dieser Stelle bitte fröhliches Pfeifen einblenden\*  
 \*drum roll\* / \*at this point please jolly whizzling fade in\*

Less frequent, but nonetheless existent, are instances of noninflected forms which are syntactically integrated, as shown in (26), (27). These challenge the analysis of noninflected verbs as independent interactive units [2, 3], classified in the same category as interjections, answer particles, emoticons and user names, and support an analysis which integrates non-inflected verb forms in the verbal paradigm.

- (26) Jetzt mal erst so \*tür aufmach\* und dann \*rausgeh\*  
 now PTCL first so \*door open<sub>noninflected</sub>\* and then \*step<sub>noninflected\_out</sub>\*  
 ”\*opening door\* now and \*stepping out\*“
- (27) dafür \*zitter\* und \*dick einmumm\*  
 instead shiver<sub>noninflected</sub> and \*thick wrap\_up\_well<sub>noninflected</sub>\*  
 ”instead shivering and wrapping myself up well“

Complex inflective constructions, on the other hand, pose a major challenge for automatic POS tagging, as they are often written as one token, sometimes (but not always) separated by space or by hyphens. Depending on the way they have been transcribed, they will either be split up into individual tokens or will be treated as single unit by the tokeniser.

In previous work [11], we have annotated complex inflective constructions using the COMMENT label. This, however, is not sufficient to encode the rich information expressed by these units. Here we expand our analysis and argue that the components of these constructions should be tokenised and annotated as individual units on the sub-token level. Figure 1 displays the syntactic structure of the inflective construction and the complex hashtag in (28), neither of which is syntactically integrated in the tweet. On the token level, the inflective construction and the hashtag are both treated as one unit. A more detailed analysis of the internal

<sup>10</sup>Inflectives (non-inflected verb forms) are a frequent stylistic means in German comics and computer-mediated communication [16].

structure of the two constructions is given on a sub-token level, where the complex inflective construction and the hashtag are split up and analysed individually.

- (28) @pillenknick Moin Hendryk , schon unterwegs ? \*Kaffeetasseheb\*  
 @username Morning Hendryk , already on your way ? \*coffee-cup-lift\*  
 #nochimmerschläfrig  
 #still-tired

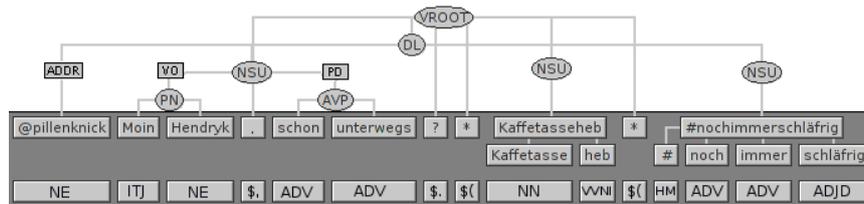


Figure 1: Analysis of complex inflectives and hashtags on the sub-token level

The syntactic analysis follows the annotation scheme of the TIGER treebank [4] as closely as possible. The DL (discourse level) node is the top node of the tweet. The user name (@pillenknick), referencing the addressee of the tweet, is marked by the new grammatical function label ADDR.<sup>11</sup> The actual tweet message (Moin Hendryk, schon unterwegs?) is governed by a NSU (non-sentential unit) node,<sup>12</sup> as are the inflective construction and the hashtag. We do not include the asterisks as part of the token, as they are not a necessary component of the inflective construction, as opposed to the # for hashtags.

Figure 2 illustrates the representation of integrated inflectives on the POS level and in the syntax.

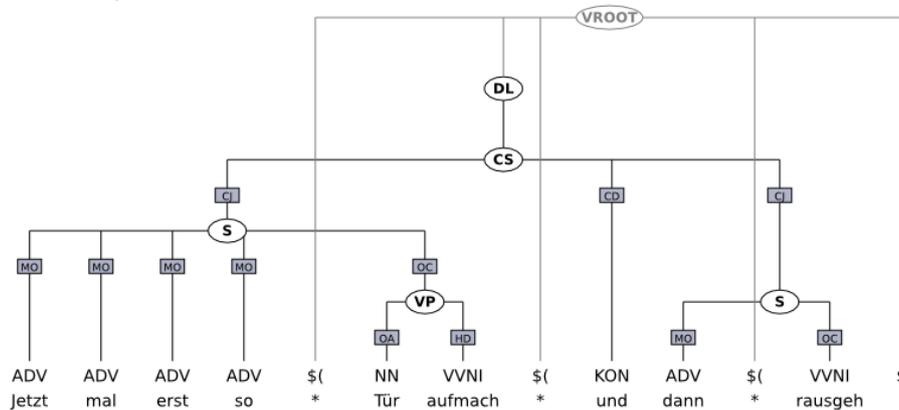


Figure 2: Analysis of integrated non-inflected verb forms

<sup>11</sup>Legend for figures 1 and 2: *Syntactic categories*: DL: discourse level, S: clause, CS: coordinated clause, NSU: non-sentential unit, PN: proper name, AVP: adverbial phrase, VP: verb phrase; *Grammatical functions*: ADDR: addressee, VO: vocative, PD: predicate, OP: prepositional object, CJ: conjunct, CD: coordinating conjunction, MO: modifier, OC: clausal object; *POS*: NE: proper name, ITJ: interjection, ADV: adverb, NN: noun, VVNI: non-inflected verb, HM: hashtag marker, \$,: comma, \$(: sentence-final punctuation, \$(: sentence-internal punctuation.

<sup>12</sup>We distinguish between sentential and non-sentential units. Sentential units do include a finite verb while NSU nodes don't.

To sum up, hashtags not only provide a semantic classification of the tweets but also allow the users to express their emotions or comment on their physical condition or the state of the world in general. They do not correspond to one particular part of speech but can take the form of any arbitrary word or construction, of sentences even, depending on the creativity of the users. We thus argue that hashtags should not be annotated with a special hashtag label but should be analysed according to their distributional properties and internal structure.

## 5 Conclusions

This paper contributes to the discussion on best practices for the syntactic analysis of non-canonical language, focusing on Twitter microtext. We first presented an annotation experiment where we tested proposals from the literature for POS annotation of tweets and compared our inter-annotator agreement to related work. While our overall inter-annotator agreement was in line with, or even higher than, what has been reported in comparable studies, our error analysis showed that especially the annotation of Twitter-specific phenomena such as hashtags and mentions causes disagreements between human annotators. We argued that the new POS tags introduced to label user names and hyperlinks do not correspond to new grammatical part-of-speech categories. Accordingly, we advocate the annotation of user names and hyperlinks as proper names.

Furthermore, we discussed the multiple functions of the @- and #-sign in Twitter, showing that POS tags like AT-MENTION or HASHTAG fall short of capturing the information encoded by these phenomena. Instead, we sketched a possible way of annotating complex non-inflected constructions and hashtags in the syntax tree, providing a coarse-grained analysis on the token level and a more detailed one on the sub-token level.

## References

- [1] T. Avontuur, I. Balemans, L. Elshof, N. van Noord, and M. van Zaanen. Developing a part-of-speech tagger for Dutch tweets. *Computational Linguistics in the Netherlands Journal*, 2:34–51, 2012.
- [2] M. Beißwenger, M. Ermakova, A. Geyken, L. Lemnitzer, and A. Storrer. A TEI schema for the representation of computer-mediated communication. *Journal of the Text Encoding Initiative*, (3):1 – 31, 2012.
- [3] M. Beißwenger, M. Ermakova, A. Geyken, L. Lemnitzer, and A. Storrer. De-RiK: A German reference corpus of computer-mediated communication. *Literary and Linguistic Computing*, 4(28):531–537, 2013.
- [4] S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith. The TIGER treebank. In *Proceedings of TLT*, Sozopol, Bulgaria, 2002.

- [5] T. Brants. Inter-annotator agreement for a german newspaper corpus. In *Proceedings of LREC*, Athens, Greece, 2000.
- [6] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N.A. Smith. Part-of-speech tagging for Twitter: annotation, features, and experiments. In *Proceedings of ACL*, Portland, Oregon, 2011.
- [7] E. Hinrichs, S. Kübler, K. Naumann, H. Telljohann, and J. Trushkina. Recent developments in linguistic annotations of the TüBa-D/Z treebank. In *Proceedings of TLT*, Tübingen, Germany, 2004.
- [8] P. Koch and W. Oesterreicher. Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36(85):15–43, 1985.
- [9] W. Ong. *Orality and literacy: The technologizing of the word*. London; New York: Methuen, 1982.
- [10] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N.A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL*, Atlanta, Georgia, 2013.
- [11] I. Rehbein. Fine-grained POS tagging of German tweets. In *Proceedings of GSCL*, Darmstadt, Germany, 2013.
- [12] I. Rehbein and S. Schalowski. Extending the STTS for the annotation of spoken language. In *Proceedings of KONVENS 2012*, Vienna, Austria, 2012.
- [13] I. Rehbein, S. Schalowski, N. Reinhold, and E. Visser. Ähm, äh... filled pauses in computer-mediated communication. Potsdam, Germany, 2012. Talk presented at the DGfS Workshop on "Modelling Non-Standardized Writing".
- [14] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: an experimental study. In *Proceedings of EMNLP*, Edinburgh, United Kingdom, 2011.
- [15] A. Schiller, S. Teufel, and C. Thielen. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, IMS-CL, University Stuttgart, Germany, 1995.
- [16] P. Schlobinski. \*knuddel – zurueckknuddel – dich ganzdollknuddel\*. Inflektive und Inflektivkonstruktionen im Deutschen. *Zeitschrift für germanistische Linguistik*, 29(2):192–218, 2001.