

Ines Rehbein and Sören Schalowski

STTS goes Kiez – Experiments on Annotating and Tagging Urban Youth Language

1 Introduction

The Stuttgart-Tübingen Tag Set (STTS) (Schiller et al., 1995) has long been established as a quasi-standard for part-of-speech (POS) tagging of German. It has been used, with minor modifications, for the annotation of three German newspaper treebanks, the NEGRA treebank (Skut et al., 1997), the TiGer treebank (Brants et al., 2002) and the TüBa-D/Z (Telljohann et al., 2004). One major drawback, however, is the lack of tags for the analysis of language phenomena from domains other than the newspaper domain. A case in point is spoken language, which displays a wide range of phenomena which do not (or only very rarely) occur in newspaper text.

The STTS, as a consequence, does not provide POS tags to capture these phenomena. As a result, other POS categories have been stretched to describe spoken language. For instance, in the Tübingen Treebank of Spoken German (TüBa-D/S) (Stegmann et al., 2000) the tag for interjections has been used to annotate filled pauses and backchannel signals like *uh*, *mhm*, adjectives like *richtig*, *gut*, *hervorragend* (right, good, excellent) when used in isolation, and for question tags. From a linguistic point of view, this practice is unsatisfactory and should be given up in favour of a more adequate description of spoken language phenomena.

In this paper, we present an extension of the STTS for the annotation of spoken language. We describe our new tagset and evaluate its adequacy in a manual annotation experiment. Furthermore, we develop a POS tagger for analysing spoken language data and evaluate its performance on spoken language transcripts as well as on a normalised version of the data.

This paper is structured as follows. In Section 2 we motivate the need for an extension of the STTS and define the new tags. Section 3 describes the data used in our experiments and presents the results of an annotation experiment. We report inter-annotator agreement on the extended tagset and make a proposal for restructuring and integrating the new tags into the STTS. In Section 4 we report on our efforts to develop a tagger for spoken language data, describing the tagging architecture and basic features used in our experiments. Section 5 focusses on adapting the tagger to spoken language, especially on addressing the out-of-vocabulary (OOV) problem of our data. We conclude and outline future work in Section 6.

POS	TiGer	TüBa-D/Z	TüBa-D/S
PTKANT	7.5	26.2	279.7
ADV	27.1	71.6	46.7
ITJ	0.4	0.0	0.0
NN	1.8	2.4	0.0
KON	0.0	2.0	0.0
TOTAL	36.8	102.2	326.4

Table 1: Distribution of *ja* (yes) in different corpora, normalised by corpus size

2 Extensions to the STTS tag set

This section describes our extensions to the STTS for the annotation of spoken language. We first motivate the need for additional POS tags for analysing spoken language data. We review related work and argue that extending an existing annotation scheme is preferable to developing a new scheme tailored towards the specific needs of spoken language. Then we present our new tags and describe their usage.

2.1 (Why) do we need new tags?

A major pitfall for the annotation of spoken language is the danger of carrying over annotation guidelines from standard written text which, at first glance, seem to be adequate for the description of spoken language, too. Only at second glance does it become obvious that what looks similar at first may not necessarily be the same.

A case in point is *ja* (yes), which in written text mostly occurs as a modal particle in the middle field, while in spoken language occurrences of *ja* in utterance-initial position constitute the more frequent type. Table 1 shows the distribution of *ja* in two German newspaper corpora, the TiGer treebank (Brants et al., 2002) and the TüBa-D/Z (Release 8) (Telljohann et al., 2004), and in a corpus of spoken dialogues, the TüBa-D/S (Stegmann et al., 2000). In TiGer and TüBa-D/Z, most instances of *ja* are in the middle field, annotated as ADV, while the utterance-initial instances in the TüBa-D/S are assigned the tag PTKANT (answer particle). Motivated by the difference in distribution, we take a closer look at these instances and observe that many of the utterance-initial cases are in fact discourse markers (Example 1).

- (1) **ja** wer bist du denn ?
 PTCL who are you then ?
 And who are you now?

Other phenomena which cannot be analysed using the STTS inventory are filled pauses, question tags and backchannel signals. In the TüBa-D/S, filled pauses have been removed from the corpus, while question tags have been analysed as interjections (Example 2), as have backchannel signals (Example 3).

- (2) es war doch Donnerstag , **ne** ?
 is was however Thursday , no ?

It was Thursday, wasn't it?

- (3) **mhm** ja das ist bei mir ganz offen
uh-huh yes this is for me totally open
Uh-huh, yes, I'm quite flexible.

We argue that these instances should not be analysed as interjections, as done in the TüBa-D/S, but should be assigned a new POS tag. In the next section, we report on related work on the annotation of word classes in spoken language corpora.

2.2 Related work

A number of spoken language corpora already exist, annotated with parts of speech. However, not much work has been done on developing or extending POS tagsets for the annotation of spoken language. Many corpora use POS tagsets originally developed for written language or make only minor changes to the tagset.

The Tübingen Treebank of Spoken German (TüBa-D/S), for instance, uses the STTS which had been developed for the annotation of written language. The spoken part of the BNC applies a tagset with around 60 tags but does not encode spoken language phenomena on the POS level. Hesitations in the BNC are not considered to be linguistic words and are thus annotated as *unclassified items*, as are non-English words, special typographical symbols and formulae. Discourse markers, on the other hand, such as backchannel signals and question tags, are subsumed under the interjection label.

The Switchboard corpus (Godfrey et al., 1992), a corpus of spontaneous conversations of English telephone bandwidth speech, follows the tagging guidelines of the Penn Treebank POS tagset, which was developed for annotating written (newspaper) text.¹ Switchboard only introduced minor changes to the original tagset. They added the BES and HVS tags to distinguish between *is* and *has* when being contracted and reduced to 's. Another extension is the XX tag used for marking partial words where the full word form can not be recovered from the context. Similarly to the BNC, different discourse markers are treated as interjections.

One example for a POS tagset specifically designed for annotating spoken language is the one developed for the Spoken Dutch Corpus (Oostdijk, 2000). The hierarchical tagset distinguishes 10 major word classes, while the whole tagset provides more than 300 fine-grained morpho-syntactic tags (Eynde et al., 2000). Despite its detail, the tagset does not encode differences between different markers of discourse but, similar to the BNC, analyses these items as interjections.

Two noteworthy exceptions to the simple re-use of schemes developed for written language are the Vienna-Oxford International Corpus of English (VOICE) (Breiteneder et al., 2006), a corpus of English as a lingua franca, and the the Christine corpus (Sampson, 2000), which is one of the first treebanks of spoken language data.

¹The Switchboard corpus does, however, provide a fine-grained annotation of disfluencies on the syntactic level, covering phenomena such as non-sentential units and restarts.

VOICE adapts the Penn Treebank POS tagset by adding 26 new categories to the tagset. Some of them describe properties of spoken discourse (e.g. discourse markers, response particles, and formulaic items like greetings), while others add non-verbal information (e.g. breathing or laughter). Other additional tags distinguish between contracted verb forms, similar to Switchboard.

The Christine corpus uses a much more fine-grained POS tagset with more than 400 morpho-syntactic tags tailored to the analysis of spoken language. The POS tags in the Christine corpus allow one to annotate discourse phenomena such as filled pauses, backchannel signals and question tags, to distinguish between different types of swearwords, to annotate formulaic expressions like greetings, or to encode onomatopoeia and forms of echoism. The tagset also distinguishes between different types of pragmatic units, such as apologies, responsiveness, and positive and negative answers.

In the next section, we present our own work on extending the STTS for the annotation of spoken language.

2.3 Extensions to the STTS for spoken language annotation

Our approach to extending the STTS is purely *data-driven*. We started annotating the data using the original STTS tagset, and only when encountering phenomena which could not be described within the STTS tagset, we introduced new tags. We tested these provisional tags on new data and refined our classification, merging classes which proved to be difficult for the human annotators. As a result, we ended up with 11 additional tags for the annotation of spoken language phenomena (Table 2).

POS	description	example	literal translation
PAUSE	<i>pause, silent</i>	so ein (-) Rapper	such a (-) rapper
PTKFILL	<i>particle, filler</i>	ich äh ich komme auch .	I er I come too
PTKINI	<i>particle in utterance-initial position</i>	ja kommst du denn auch ?	PART come you then too
PTKRZ	<i>backchannel signal</i>	A: ich komme auch . B: hm-hm .	A: I come too B: uh-huh
PTKQU	<i>question particle</i>	du kommst auch . Ne ?	you come too . no ?
PTKONO	<i>onomatopoeia</i>	das Lied ging so lalala .	the song went so lalala
PTKPH	<i>placeholder particle</i>	er hat dings hier .	he has thingy here
VVNI	<i>uninflected verb</i>	seufz	sigh
XYB	<i>unfinished word, interruption</i>	ich ko #	I co #
XYU	<i>uninterpretable</i>	(unverständlich) #	(uninterpretable) #
\$#	<i>unfinished utterance</i>	ich ko #	I co #

Table 2: Additional POS tags for spoken language data

2.3.1 Hesitations

The first two tags encode silent pauses and filled pauses. The PAUSE tag is used for silent (unfilled) pauses which can occur at any position in the utterance.

- (4) das ist irgend so ein (-) Rapper
that is some such a rapper
That is some rapper.

The PTKFILL tag is used for filled pauses which can occur at any position in the utterance.

- (5) das ist irgend so ein äh Rapper
that is some such a uh rapper
That is some uh rapper.

2.3.2 Discourse particles

The following tags are used for the annotation of discourse particles. We use the term *discourse particles* in a theory-neutral sense as an umbrella term for a variety of particles and discourse markers frequently used in spoken language.

The PTKINI tag is assigned to particles such as *ja* (yes), *na* (there, well) when used as a discourse marker in an utterance-initial position. In contrast to interjections, these particles do not carry stress. They have been described in the literature as *Eröffnungssignale* (opening signals) (Schwitalla, 1976) or *Gliederungssignale* (discourse structuring signals) in utterance-initial position (Schwitalla, 2006), or as discourse markers in the pre-prefield (Auer and Günthner, 2005).

- (6) **ja** wer bist du denn ?
PTCL who are you then ?
And who are you now?

Please note that most occurrences of *ja* (yes) in the middle field are modal particles (Example 7) which are assigned the ADV label (adverb) in the German treebanks. Occurrences of *ja* in utterance-initial position, on the other hand, are discourse markers and thus should be treated differently (also see Meer (2009) for a discussion on the different word classes of *ja*).

- (7) die hat **ja** auch nicht funktioniert .
that one has PTCL also not worked .
That one didn't work, either.

The PTKRZ tag is used for backchannel signals. We define backchannel signals as plain, non-emotional reactions of the recipient to signal the speaker that the utterance has been received and understood.

- (8) A: stell dir das mal vor !
 A: imagine you this PTCL VERB PTCL !
 Imagine that !
- (9) B: **m-hm**
 B: uh-huh

Preliminary annotation experiments showed a very low inter-annotator agreement for the distinction between answer particles and backchannel signals for *ja*. There is, in fact, an overlap of meaning which makes a clear distinction between the function of *ja* as an answer particle and a backchannel signal infeasible. To support consistency of annotation, we always label *ja* as answer particle and not as backchannel signal.

The PTKQU tag is used for question tags such as *ne* (no), *gell* (right) or *wa* (what) added to the end of a positive or negative statement. Please note that we do not annotate adjectives like *okay*, *richtig* (okay, right), interrogative pronouns like *was* (what) or conjunctions like *oder* (or) as PTKQU, as both classes show distributional differences. Instances of *okay*, *richtig*, *was*, *und*, *oder* in the context of a question are still annotated as adjectives, interrogative pronouns or conjunctions, respectively.

- (10) wir treffen uns am Kino , **ne** ?
 we meet REFL at the cinema , no ?
 We'll meet at the cinema, right ?
- Du kommst auch , **wa** ?
 You come too , what ?
 You'll come too, right?

2.3.3 Other particles

The PTKONO tag is used for labelling onomatopoeia and forms of echoism.

- (11) das Lied ging so **lalalala**
 The song went like lalalala
- (12) **eieieieia** !
- (13) **bam** , **bam** , **bam** !
- (14) interessant , **bla bla** .
 interesting , bla bla .

The PTKPH tag is used as a placeholder when the correct word class cannot be inferred from the context. Example (15), for instance, has a number of possible readings. In (a), the correct POS tag would be noun (NN), while in (b) we would assign a past participle (VVPP) tag. The placeholder might also stand for a whole VP, as in (c).

- (15) er hat **dings** hier .
 he has thingy here .
- a. er hat MP3-Player_{nn} hier .
 he has MP3 player here .
- b. er hat gewonnen_{vvp} hier .
 he has won here .
- c. er hat (Schuhe gekauft)_{vp} hier .
 he has shoes bought here .

2.3.4 Uninflected verb forms

We use the tag *VVNI* to annotate non-inflected verb forms (Teuber, 1998). Non-inflected auxiliaries (*VANI*) and modal verbs (*VMNI*) are also possible forms but very rarely occur in spoken language. They do, however, occur in computer-mediated communication (CMC).

- (16) ich muss noch putzen . **seufz** !
 I must still clean . sigh !
 I still have to clean. Sigh!
- (17) gleich haben wir Mathe . **gäh** !
 soon have we math . yawn !
 We have math right now. Yawn!

2.3.5 Non-words

The STTS provides the *XY* tag for the annotation of non-words. We add two new subclasses to distinguish between different types of non-words.

1. uninterpretable material (*XYU*)
2. unfinished words (*XYB*)
3. other (*XY*)

The *XYU* tag is used for lexical material which is uninterpretable, mostly because of poor audio quality of the speech recordings or because of code-switching.²

- (18) wir waren gestern bei (**fremdsprachlich**).
 we were yesterday at (FOREIGN).
 Yesterday we've been at (FOREIGN).

The *XYB* tag is used for abandoned words.

²For foreign language material in the data which can be understood and transcribed we use the *FM* tag provided by the STTS.

- (19) ich habe **gest** # heute komme ich nicht .
I have yest # today come I not .
I have yest- # I won't come today.

The XY tag is used for all non-words which do not fit one of the categories above. This category is more or less consistent with the XY category in the STTS where it is used for non-words including special symbols.

2.3.6 Punctuation

The \$# tag is a new punctuation tag used to mark interrupted or abandoned utterances. These can (but do not necessarily) include unfinished words, as in Example (20).

- (20) sie war gest #
she was yest #

2.3.7 Extensions to the STTS – Conclusion

The corpora presented in Section 2.2 made different decisions regarding the question what kind of information should be encoded on the POS level. Some of them try to restrict the tagset to word classes which can be defined purely on a grammatical level (TüBa-D/S, BNC, Switchboard, Spoken Dutch Corpus), others choose to also include rich pragmatic information (VOICE, Christine). While it is hard to stick to a purely grammatical distinction – the STTS, for instance, uses a mix of grammatical, distributional and semantic criteria for defining different word classes – the latter approach is not uncontroversial, either. Pragmatic categories are often vague and ill-defined, thus compromising the consistency of the annotations. It can also be argued that they provide a very different type of information which should not be encoded on the word class level.

While that point is well taken, we would still like to include pragmatic information, which is highly relevant for the analysis of discourse, in the corpus. We consider the annotation layers of the corpus not as the final product but as a database which allows us to generate different views on the data (which would correspond to different corpus versions of the same data, one subsuming all discourse particles under the *interjection* label, another one also including the pragmatic tags on the same level or projecting those to a new annotation layer). Our reasons for encoding pragmatic information on the POS level are mostly practical ones. This way of proceeding allows for swift annotation without the need for a second pass over the data, it results in a more compressed, thus more clearly arranged presentation of the data (whereas adding yet another corpus layer would give us a more confusing view), and, finally, it also facilitates corpus queries.

3 Annotation experiments

This section reports on an annotation experiment with human annotators using the extended tagset. We describe the data we used in the experiments and report numbers for inter-annotator agreement (IAA) between the annotators. Based on a detailed error analysis for the new POS tags we present our proposal for integrating the new tags into the STTS.

3.1 Data: KiDKo – The Kiezdeutsch-Korpus

The data we use in our experiments is taken from the Kiezdeutsch-Korpus (KiDKo) (Wiese et al., 2012). Kiezdeutsch (*hood German*) is a variety of German used in informal peer-group communication, spoken by adolescents from multilingual urban neighbourhoods.

The data was collected in the first phase of project B6 “Grammatical reduction and information structural preferences in a contact variety of German: Kiezdeutsch” as part of the SFB (Collaborative Research Centre) 632 “Information Structure” in Potsdam. It contains spontaneous peer-group dialogues of adolescents from multiethnic Berlin-Kreuzberg (around 48 hours of recordings) and a supplementary corpus with adolescent speakers from monoethnic Berlin-Hellersdorf (around 18 hours of recordings). The current version of the corpus includes the audio signals aligned with transcriptions. The data was transcribed using an adapted version of the transcription inventory GAT basic (Selting et al., 1998), often referred to as minimal GAT, including information on primary accent and pauses. Additional annotation layers (POS, Chunking, Topological Fields) are work in progress.³

The transcription scheme has an orthographic basis but, in order to enable investigations of prosodic characteristics of the data, it also tries to closely capture the pronunciation, including pauses, and encodes disfluencies and primary accents. In addition, we are adding a level of orthographic normalisation where non-canonical pronunciations and capitalisation are reduced to standard German spelling. This annotation layer enables us to use standard NLP tools for semi-automatic annotation.⁴ It also increases the usability of the corpus as it allows one to find all pronunciation variants of a particular lexeme. The normalisation is done in a semi-automatic way. We copy the text from the transcription layer, automatically correcting frequent deviations from the orthographic norm based on dictionaries and frequency lists. The remaining changes are carried out manually during the transcription process.

Figure 1 shows an example transcript from the KiDKo, displaying the transcription and the normalisation layer, the POS tags and the layers for non-verbal information. Uppercase letters on the transcription layer mark the main accent of the utterance. The equals sign is used to encode the tight continuation of a word form with a following

³The first release of KiDKo is scheduled for spring 2014 and will include the transcribed data as well as the normalisation and POS annotation layers.

⁴Please note that we still have to adapt these tools to our data. However, without the normalisation the manual effort to correct these tags would be much higher.

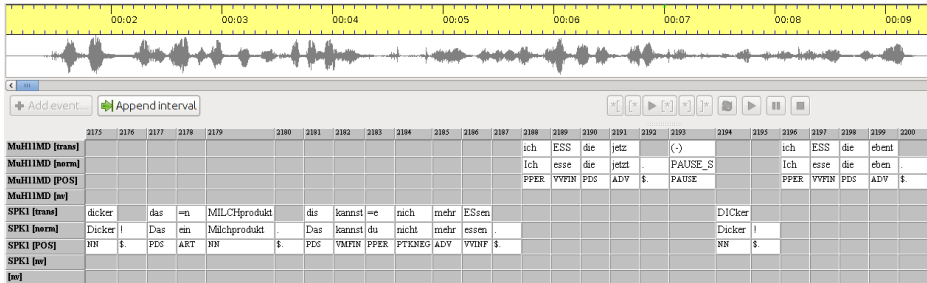


Figure 1: Screenshot KiDKo sample of a short dialogue between two speakers (MuH11MD, SPK1) in EXMARaLDA (transcription, normalisation, POS and non-verbal layer)

form, where one of the two forms (or both of them) is reduced (e.g. das =n “this a”, kannst =e “can you”).

We would like to emphasise that linguistic annotations not only provide a description but also an interpretation of the data. This is especially true for the annotation of learner data, where the formulation of target hypotheses has been discussed as a way to deal with the ambiguity inherent to a learner’s utterances (Hirschmann et al., 2007; Reznicek et al., 2010). When annotating informal spoken language, we encounter similar problems. Adding an orthographic normalisation to the transcription might be seen as a ‘poor man’s target hypothesis’ where decisions made during the annotation become more transparent.

3.2 Inter-annotator agreement

Inter-annotator agreement for human coders on the core STTS is quite high. For instance, Rehbein et al. (2012) report a percentage agreement of 97.9% and a Fleiss’ κ of 0.978 for two human annotators on the target hypotheses of essays written by advanced second language learners of German.

In a preliminary annotation experiment with three human annotators, we obtained a percentage agreement of 96.5% and a κ of 0.975 on a small test set (3,415 tokens) from the KiDKo, using our extended tagset. This shows that the extended tagset does not result in a decrease in accuracy for the manual annotation.

However, due to the small size of the test set, some of the new tags only occurred infrequently in the sample. To provide a more meaningful evaluation for the new tags, we created a new test set, focussing only on the discourse particles *answer particles*, *backchannel signals*, *question tags*, *fillers*, *utterance-initial particles* (PTKANT, PTKRZ, PTKQU, PTKFILL, PTKINI) and on *onomatopoeia* and *placeholders* (PTKONO,

POS	freq.	# agr.	% agr.
PTKANT	903	809	89.59
PTKQU	296	255	86.15
PTKFILL	126	112	88.89
PTKINI	121	116	95.87
PTKONO	61	55	90.16
PTKRZ	33	15	45.45
PTKPH	13	8	61.54
avg.	1553	1370	88.22

Table 3: IAA for two human annotators on the discourse particles

PTKPH).⁵ We took a subpart of the corpus with 39,583 tokens which had already been annotated with POS by one annotator. A second annotator then assigned POS tags to all instances which she considered to be one of the discourse particles listed above. Candidate instances for the second annotation were identified using heuristics based on automatically assigned tags by two versions of the TreeTagger, one using the tagging model trained on the TiGer corpus and the second one using a tagging model adapted to the new tags.⁶ The accuracy of the two tagging models on the KiDKo data is not very high. The two taggers do, however, produce systematic errors on the new domain which allows us to detect instances of discourse particles without having to look at too many tokens. In the evaluation, we only include those instances which had been assigned one of the discourse particle tags by at least one of the annotators. Table 3 shows detailed results for these tags.

For all particle tags we observe an agreement which is below the average agreement on the whole tagset. This is not surprising, as these tags do encode pragmatic information and are thus much harder to define and operationalise than most of the other POS tags.

For some categories, we obtained an acceptable agreement of close to or over 90% (PTKINI, PTKANT, PTKFILL, PTKONO). For the question tags (PTKQU), many disagreements were caused by one annotator assigning the PTKQU tag to conjunctions like *oder*, *und* (or, and) when used in the context of a question, while the second annotator assigned the KON tag to these instances (Example 21), as intended by the guidelines. This problem can easily be solved by revising the guidelines and making explicit which tokens should be interpreted as a question tag and which should not.

- (21) du musst au wir müssen AUFhören **oder** (transcript)
 Du musst au wir müssen aufhören . Oder ? (normalisation)
 you have to sto we have to stop . or (literal translation)

⁵Silent pauses and \$# have not been included in the testset because they are not ambiguous. The XYB/XYU tags do not encode linguistically interpretable categories but should be considered as a technical device to deal with material which, partly caused by low audio quality and partly caused by phenomena of social interaction, otherwise could not be analysed.

⁶The adapted model was also trained on the TiGer corpus but uses an extended dictionary which includes word-POS pairs for the most frequent tokens in the KiDKo data.

You have to sto- we have to stop, right? (free translation)

Only low agreement could be achieved on the placeholder particles. Here the annotators disagreed on whether instances of the following kind are ambiguous between different POS tags or not (Example 22). According to the guidelines, Example 22 should be analysed as a placeholder because the placeholder slot could be filled with a noun (Example 22 a) or a verb (Example 22 b). Unfortunately, the annotators are not always aware of these ambiguities but tend to settle on the most probable interpretation of the utterance, thus overlooking other possible readings.

- (22) wenn ich das hier äh (-) **DINGS** (-) äh a (-) wenn ich das ANschalte
Wenn ich das hier äh **dings** äh a Wenn ich das anschalte
when I this here uh thingy uh a when I this turn on
- a. Wenn ich das hier äh **Schalter** äh a (-)
when I this here uh button uh a
- b. Wenn ich das hier äh **anschalte** äh a (-)
when I this here uh turn on uh a

The hardest task for the human annotators was the distinction between answer particles, backchannel signals and fillers. For illustration, consider Example 23, where one annotator interpreted the first particle, *hm*, as a filler (PTKFILL) and the second one, *'hmhm*,⁷ as an answer particle, while the second annotator analysed both particles as backchannel signals (PTKRZ). In Example 24, on the other hand, annotator one interpreted the token as a backchannel signal while annotator two annotated it as an answer particle (PTKANT).

- (23) A: schmeckt ja dann bestimmt SCHEIße (-) **hm** **'hmhm**
A: Schmeckt ja dann bestimmt scheiße . **Hm** . **Hm-hm** .
A: tastes PTC then surely shit hm m-hm
A: This surely will taste like shit. Hm. M-hm.
- (24) B: als wenn man da DRÖgn vertickt aufm schulhof (-) A: **hm**
B: Als wenn man da Drogen vertickt aufm Schulhof . A: **Mh**
B: as if one there drugs sells at the schoolyard A: mh
B: As if one would sell drugs on the schoolyard. A: Mh .

It is not clear whether these distinctions can be operationalised sufficiently to enable a reliable annotation. One might ask whether it is advisable to encode pragmatic differences like those in a part-of-speech tagset or whether these fine-grained annotations should be transferred to a separate layer of annotation, and should be subsumed under the answer particles on the part-of-speech level.

⁷The apostrophe indicates a glottal stop.

coarse	fine
	DMANT answer particles (PTKANT)
	DMITJ interjections (ITJ)
DM	DMQU question particles
discourse markers	DMRZ backchannel signals
	DMFILL filler
	DMINI utterance-initial discourse particle
PTKONO	onomatopoeia
PTKPH	placeholder
VANI/VMNI/VVNI	uninflected verbs

Table 4: Possible integration of the new tags in the STTS

As a compromise, we propose the following classification of the new tags, shown in Table 4. A coarse-grained POS tag for discourse markers could ensure a reliable, consistent annotation, while a more fine-grained classification can be used when a more detailed analysis is wanted. The DM tag would now comprise the former STTS tags for answer particles (PTKANT) and interjections (ITJ) as well as question particles, backchannel signals, fillers and utterance-initial discourse particles. In addition to the STTS tags for separable verb particles (PTKVZ), the particle *zu* with an infinite verb form (PTKZU) and a particle with an adjective or adverb (PTKA), we now have the placeholder particle (PTKPH) and the particle for onomatopoeia and forms of echoism (PTKONO).⁸ The non-inflected verb forms (VANI/VMNI/VVNI) are part of the STTS verb paradigm, as indicated by the prefix VA/VM/VV.

4 Developing a POS tagger for Kiezdeutsch

While automatic POS tagging of canonical, written text from the newspaper domain might appear to be a solved problem with accuracies in the high nineties (Schmid, 1994, 1995; Schmid and Laws, 2008), a very different picture emerges when looking at text from other domains. Applying a tagger trained on newspaper text to spoken language data or to user-generated content from the web will result in a substantial decrease in accuracy. The use of informal language, creative inventions of new words and a high number of variants for existing word forms in combination with a non-canonical syntax result in data sparseness and causes problems for automatic processing. For spoken language, disfluencies such as hesitations, filled pauses, repeated material or repairs further add to the problem.

This section describes our efforts to develop a POS tagger for spontaneous multiparty dialogues of Kiezdeutsch. The data used in our experiments includes 18 different transcripts, where each transcript has between two and seven speakers. Table 14 (Appendix) shows the distribution of POS tags in the manually annotated data from the

⁸It is open to discussion whether onomatopoeia and placeholders should be integrated as particles or as subordinate XY elements.

corpus	tagging ambiguity
KiDKo (normalised)	1.10
KiDKo (transcripts)	1.11
TiGer	1.03

Table 5: Tagging ambiguity for KiDKo (normalised and transcribed layer) and for an equally-sized portion of the TiGer corpus

KiDKo (training/development/test sets, normalised; 28,827 tokens) and, for comparison, in a test set of the same size from the TiGer treebank. As expected, we can observe crucial differences between the data sets. Average sentence length in the newspaper corpus is much longer than the average length of utterances in KiDKo, as indicated by the higher number of sentence delimiters (\$., \$#). The exact numbers, however, should be interpreted with care, as the question of how to segment spoken language is by no means clear.⁹

Our guidelines for segmentation are motivated by our goal of maintaining interoperability and comparability with corpora of written language on sentential utterances in the data. We follow the terminology of Fernandez and Ginzburg (2002) and define *sentential utterances* as utterances containing a finite verb, while we call utterances without a finite verb *non-sentential utterances*. We thus base our unit of analysis on structural properties of the utterance and, if not sufficient, also include functional aspects of the utterance as criteria for segmentation. The latter results in what we call the principle of the *smallest possible unit*, meaning that when in doubt whether to merge lexical material into one or more units, we consider the speech act type and discourse function of the segments. Example (25) illustrates this by showing an utterance including an imperative (*Speak German!*) and a check question (*Okay?*). It would be perfectly possible to include both in the same unit, separated by a comma. However, as both reflect different speech acts, we choose to annotate them as separate units of analysis.

- (25) Rede auf Deutsch ! Okay ?
 Speak on German ! Okay ?
 Speak German! Okay?

Other striking differences between KiDKo and TiGer include the higher number of attributive adjectives, adpositions, articles, nouns and proper names in TiGer, while in KiDKo we observe a higher frequency of adverbs, personal and demonstrative pronouns, finite auxiliaries and imperatives and, of course, of answer particles.

The tagging ambiguity (number of tags per token) for the KiDKo is 1.10 for the normalised transcripts and 1.11 for the original transcripts (Table 5). For comparison, the tagging ambiguity for an equally-sized portion from the TiGer treebank is 1.03.

⁹See, e.g., Crookes (1990); Foster et al. (2000) for a discussion on the adequate unit of analysis for investigations of spoken language.

We divided the manually annotated data into a training set, a development set and a test set. The split of the data was done as follows. First we created 10 bins. Then we processed the data utterance-wise and put the first utterance in bin1, the second utterance in bin2, and so forth. As a result, we ended up with three bins holding 475 utterances each and seven bins with 474 utterances each. From this, we took the first 1,500 of the 4,743 utterances for the development set (9,210 tokens) and the next 1,500 utterances (8,935 tokens) for the test set. The remaining 1,743 utterances (10,682 tokens) were used as training data.

The transcribed version of the data has fewer tokens than the normalised version because the transcripts do not include punctuation. For the transcripts, the development set includes 7,059 tokens, the test set 6,827 tokens and the training set 8,231 tokens. Unless stated otherwise, all results reported throughout the paper are on the development set.

Before developing our own tagger, we want to establish a baseline by testing how well a state-of-the-art tagger for German, trained on newspaper text, performs on the spoken language data.

4.1 Baseline

For our baseline we use the TreeTagger (Schmid, 1994, 1995), a probabilistic POS tagger using Markov models and decision trees. We use the tagger and the German tagging model out-of-the-box with no adaptations and apply it to the spoken language data.

Preparing the input for the tagger is not straightforward, as we have multi-party dialogues with overlapping speech events. In this work we pay no attention to the temporal ordering of the utterances but use randomised sets as training/development/test data (see Section 4 above). Proceeding like this means that we lose important context information, while a more sophisticated way of presenting the data to the tagger might improve results. We will explore this in future work.

	accuracy			
	<i>transcript</i>		<i>normalised</i>	
	<i>dev</i>	<i>test</i>	<i>dev</i>	<i>test</i>
<i>baseline 1: original TreeTagger</i>				
extended tagset	42.54	42.48	74.53	73.67
core STTS tags only	51.48	51.71	86.03	85.29
<i>baseline 2: re-trained TreeTagger</i>				
extended tagset	58.42	59.90	91.76	91.56
core STTS tags only	53.49	55.28	92.22	92.08

Table 6: Baseline results for the TreeTagger on spoken language data for the original transcripts and for the normalised version of the transcripts

Table 6 shows the baseline results for the TreeTagger on the transcribed dialogues as well as on a normalised version of the development and test set. In the first row results are given for all tags in the extended tagset. Please note that this setting is rather unfair as many of the POS tags are not included in the original STTS tag set and are thus unknown to the tagger. The second row presents results on those POS tags which are included in the STTS tag set. Results show the importance of the manual normalisation of the transcripts. For the extended tagset as well as for the core STTS tags, we achieve an accuracy of more than 30% higher than on the original transcripts.

The second baseline shows results for the TreeTagger which was re-trained on the KiDKo training set. Results show that even a small amount of manually annotated data is enough to obtain a substantial increase in accuracy both for the transcripts as well as for the normalised version of the data. The better results on the extended tagset as compared to the STTS-only setting can be explained by the PAUSE tag, which is unambiguous and occurs with a high frequency in the data.

Please note that the accuracy on the original STTS tags even on the normalised transcripts is still substantially lower than the one obtained on in-domain data from German newspaper text where we can expect results in the range of 96 to 98%.

4.2 Base tagger and feature tuning

We develop our tagger based on Conditional Random Fields (CRF) (Lafferty et al., 2001), using the CRFsuite package¹⁰ (Okazaki, 2007) which provides a fast implementation of CRF for sequence labelling.

This section describes the features used in our experiments. We train the tagger on a small feature set, using features like word form, word length, or the number of digits in a word (see Table 7 for the whole feature set). In addition, we use prefix/suffix features (the first/last n characters of the input word form) as well as feature templates which generate new features of word ngrams where the input word form is combined with preceding and following word forms. Example 26 illustrates how these templates work. For instance, for the third token in (26), *irgend*, our templates extract the features in Table 8.

- (26) Das ist irgend so ein äh Rapper . (normalisation)
 this is some such a uh rapper .
 This is some uh rapper.

Table 9 (01a) presents results for the different settings for prefix/suffix size, starting from 4 up to 10. Results show a slight increase with larger prefix/suffix sizes. The differences between prefix/suffix sizes of 5 to 10, however, are not statistically significant.

As the transcripts include uppercase letters for marking the main accent of an utterance, we run a further experiment where we transform all word forms in the

¹⁰We run all our experiments with the default parameter setting (1st-order Markov CRF with dyad features, training method: limited memory BFGS).

feature	description
wrd	word form
len	word length
cap	is the word form capitalised? $\{0, 1\}$
anonym	number of X in word form
upper	number of upper case in wrd
digit	number of digits in wrd
<i>prefix/suffix features</i>	
pre N	prefix: first N characters of word form (from 1 to N)
suf N	suffix: last N characters of word form (from 1 to N)
<i>ngram features</i>	
ngrams	different ngram combinations
2grams	adjacent word forms: $w_{-2}w_{-1}$, $w_{-1}w_0$, w_0w_1 , w_1w_2 , context of w_0 : w_0w_{-1} , ..., w_0w_{-9} , w_0w_1 , ..., w_0w_9
3grams	$w_{-2}w_{-1}w_0$, $w_{-1}w_0w_1$, $w_0w_1w_2$
4grams	$w_{-2}w_{-1}w_0w_1$, $w_{-1}w_0w_1w_2$
5grams	$w_{-2}w_{-1}w_0w_1w_2$

Table 7: Feature set used in our experiments

data to lowercase letters. We expect that results for the transcripts improve while the accuracy on the normalised data should go down.

Table 9 (01b) shows results for the lowercase setting. We observe a significant improvement for the transcribed version of the data (two-sided McNemar test, $p < 0.0001$). To our surprise, the accuracy on the normalised transcripts also shows a slight increase, which again is statistically significant. As the difference between the prefix/suffix sizes of 5 to 10 is not statistically significant, we decided to run all further experiments with a size of 7.

5 Tagger adaptation

This section presents two methods for domain adaptation, both addressing the out-of-vocabulary (OOV) problem in the data. The first approach uses Latent Dirichlet Allocation (LDA) word clusters learned on unannotated data from the social media, namely from Twitter¹¹ microtext. The second approach relies on knowledge learned from a huge corpus of out-of-domain data, automatically annotated with POS tags. The first approach is motivated by the assumption that computer-mediated communication (CMC) shares some of the properties of face-to-face communication (see, e.g., Biber and Conrad (2009), Chapter 7 for a discussion of similarities and differences of CMC and face-to-face conversation). The second approach uses data from the same domain as the training data, but aims at improving tagging performance on the target domain by reducing the number of unknown words in the data.

¹¹<https://de.twitter.com>

feature	value
w_{-2}	das
w_{-1}	ist
w_0	irgend
w_1	so
w_2	ein
$w_{0,-2}$	irgend Das
$w_{0,-1}$	irgend ist
$w_{0,1}$	irgend so
$w_{0,2}$	irgend ein
$w_{0,3}$	irgend äh
$w_{0,4}$	irgend Rapper
$w_{0,5}$	irgend .
$w_{-2,-1,0}$	Das ist irgend
$w_{-1,0,1}$	ist irgend so
$w_{-2,-1}$	Das ist
$w_{-1,0}$	ist irgend
$w_{0,1}$	irgend so
$w_{1,2}$	so ein
$w_{0,1,2}$	irgend so ein
$w_{-2,-1,0,1}$	Das ist irgend so
$w_{-1,0,1,2}$	ist irgend so ein
$w_{-2,-1,0,1,2}$	Das ist irgend so ein

Table 8: Additional features extracted by the templates for *irgend*, Example 26

5.1 Tagger adaptation with LDA word clusters

Word clustering has been used for unsupervised and semi-supervised POS tagging, with considerable success (Biemann, 2006; Søgaard, 2010; Chrupała, 2011; Owoputi et al., 2012). For tagging English Twitter data, Owoputi et al. (2012) apply Brown clustering, a hierarchical word clustering algorithm, to the unlabelled tweets. During clustering, each word is assigned a binary tree path. Prefixes of these tree paths are then used as new features for the tagger.

Chrupała (2011) proposes an alternative to Brown clustering, using LDA. LDA has two important advantages over Brown clustering. First, the LDA clustering approach is much more efficient in terms of training time. Second, LDA clustering produces soft, probabilistic word classes instead of the hard classes generated by the Brown algorithm, thus allowing one word to belong to more than one cluster. Chrupała (2011, 2012) shows that the LDA approach outperforms Brown clustering on many NLP tasks. We

EXP	features	trans.	norm.
01a	pre/suf 4	83.89	93.45
	pre/suf 5	84.29	93.59
	pre/suf 6	84.51	93.59
	pre/suf 7	84.63	93.60
	pre/suf 8	84.68	93.67
	pre/suf 9	84.75	93.71
	pre/suf 10	84.78	93.71
01b	pre/suf 4 lc	86.31	93.52
	pre/suf 5 lc	86.67	93.65
	pre/suf 6 lc	87.19	93.87
	pre/suf 7 lc	87.23	93.88
	pre/suf 8 lc	87.26	93.87
	pre/suf 9 lc	87.20	93.87
	pre/suf 10 lc	87.25	93.80

Table 9: Results for different feature settings: varying prefix/suffix sizes (01a,b), lowercase word forms (01b)

thus apply the LDA clustering approach using the software of Chrupala (2011)¹² to an unlabelled corpus of Twitter data.

We decided to use Twitter data because it is freely accessible in a digitised format, it provides highly informal communication and includes many phenomena of spoken language, like fillers and repairs (27a), interjections and non-canonical word order (27b) as well as verb-less utterances (27c). While coming from a medially written register, Twitter data can in many respects be considered conceptually oral.¹³

- (27) a. Find ich nicht **gut** ... **äh schlimm** !
find I not good ... uh bad !
I don't think it's good ... uh bad!
- b. @BrandyShaloo **ah** OK **weil** **ich bin** noch nicht soweit ;)
@BrandyShaloo ah ok because I am still not ready ;)
@BrandyShaloo ah ok, because I'm not ready yet ;)
- c. Nächste Woche dann ich wieder mit voller Dröhnung .
next week then I again with full thundering .
Next week it's me again doing the full monty

¹²<https://bitbucket.org/gchrupala/lda-wordclass/>

¹³For a model of medial and conceptual orality and literacy see Koch and Oesterreicher (1985), and Dürscheid (2003) for an extension of the model to new forms of communication such as email, text messages or chat.

frequency	word form	POS
410873	einen	ART
16550	einen	PIS
8679	einen	ADJA
438	einen	NN
160	einen	VVFIN
144	einen	VVINF

Table 10: Entries for *einen* in the automatically created HGC dictionary (ART: determiner; PIS: indefinite pronoun, substitutive; ADJA: adjective, attributive; NN: noun; VVFIN verb, finite; VVINF: verb, infinite)

The Twitter data was collected from Twitter over a time period from July 2012 to February 2013. We used the Python Tweepy module¹⁴ as an interface to the Twitter Search API¹⁵ where we set the API language parameter to German and used the geocode parameter to define positions in 48 different urban and rural regions all over Germany from where we harvested our data.¹⁶ We ended up with a corpus of 12,782,097 tweets with a unique id.

Before clustering, we normalise the data. Instead of word forms, we use lemmas automatically generated by the TreeTagger (for unknown lemmas, we fall back to the word form). Our corpus contains 204,036,829 tokens and is much smaller than the one used in Owoputi et al. (2012) which includes 847,000,000 tokens of Twitter microtext.

We test different settings for LDA, setting the threshold for the minimum number of occurrences for each word to be clustered to 8, 10, 12 and 20, and induce clusters with 50 and 100 classes. Results for 50 and 100 classes are in the same range but slightly higher for 50 classes. We thus keep this setting and only report results for inducing 50 word classes.

5.2 Tagger adaptation with an automatically extracted dictionary

In our second approach to domain adaptation we stack the tagger with features extracted from an automatically created dictionary. The dictionary was harvested from the Huge German Corpus (HGC) (Fitschen, 2004), a collection of newspaper corpora from the 1990s with more than 204 billion tokens. We automatically POS tagged the HGC using the TreeTagger (Schmid, 1995). For each word form, we add the five most frequently assigned POS tags as new features. To reduce noise, we only included word POS pairs with a POS which had been predicted at least 10 times for this particular word form. As an example, consider the word form *einen* (Table 10). For *einen*, our automatically

¹⁴<http://pythonhosted.org/tweepy/html>

¹⁵<https://dev.twitter.com/docs/api/1/get/search>

¹⁶The reader should be aware that these parameters are not flawless and should be considered as an approximation rather than the plain truth.

method	transcription	normalised
Base tagger (lowercase)	87.23	93.88
<i>domain adaptation 1: LDA</i>		
LDA 50-6	88.66	95.02
LDA 50-8	88.76	94.94
LDA 50-10	88.95	95.02
LDA 50-12	88.96	95.01
LDA 50-20	88.77	94.93
<i>domain adaptation 2: HGC</i>		
HGC	90.66	95.86

Table 11: Results for LDA word clusters (trans.: lc, without cap, upper ; norm.: lc, but with features cap, upper) and for the HGC dictionary on the development set

created dictionary lists the entries in Table 10. We use the first 5 tags, ART, PIS, ADJA, NN and VVFIN, as additional features for the tagger.

5.3 Tagger adaptation – results

Table 11 presents results for the different domain adaptation methods. For convenience, we also repeat the results from Section 4.2 for our base tagger (prefix/suffix size 7).

The results for the different thresholds for the LDA word clustering approach are very close. We obtain an improvement over the base tagger of close to 2% on the transcripts and around 1% on the normalised data. Using the dictionary features from the HGC, we achieve an increase in accuracy of more than 3% on the transcripts and of nearly 2% on the normalised version of the data.

5.4 Error analysis

To find out whether the two different methods address the same problem, we analyse the most frequent errors for each setting.

Figure 2 shows the impact of the different settings on the predictions for those tags where our base tagger made the most mistakes on the transcripts (nouns: NN, adjectives: ADJD/ADJA, adverbs: ADV, finite/infinite/imperative verbs: VVFIN/VVINF/VVIMP, indefinite substitutive pronouns: PIS, foreign material: FM, verb particles: PTKVZ, demonstrative pronouns: PDS, interjections: ITJ, unfinished words: XYB, proper names: NE, determiners: ART). We see a substantial improvement for the LDA clustering approach on most POS tags. For interjections and proper names, however, the clusters do result in a higher error rate as compared to the base tagger.

The features from the HGC improve the tagging accuracy for all tags over the base tagger, and also lead to an improvement over the LDA approach on most tags. Exceptions are foreign material (FM), unfinished words (XYB) and determiners (ART).

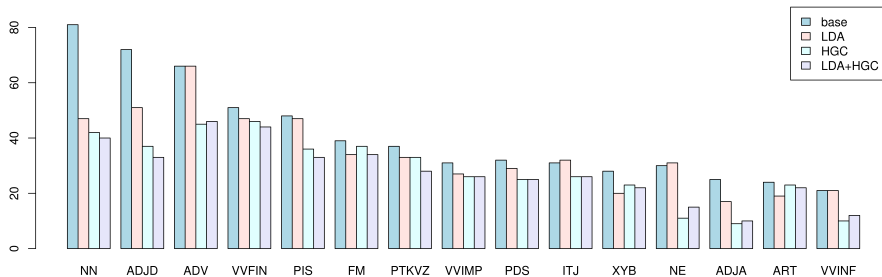


Figure 2: Error reduction for individual POS tags for the most frequent error types on the transcripts

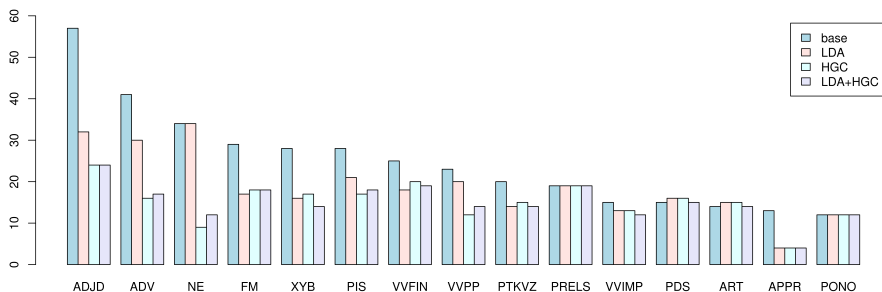


Figure 3: Error reduction for individual POS tags for the most frequent error types on the normalised data

The increased error rate for determiners can be easily explained by the different distribution of ART and PDS, the tag most commonly confused with ART, in the different resources. In the HGC, the ratio between ART and PDS, according to the TreeTagger predictions, is 33:1 while in the KiDKo we have slightly more demonstrative pronouns than determiners. This means that the features from the HGC are biased towards predicting a determiner, which has a negative impact on the accuracy of the ART and PDS tags on the KiDKo data. Additionally, determiners in the transcripts often show a deviating form like *ne* (eine; a) or *=s* (das; the) which causes the tagger to confuse it with an answer particle in the first case and with a personal pronoun in the latter case.

<i>transcription</i>						
POS	IAA (%)	freq.	base	LDA	HGC	LDA+HGC
PTKANT	89.59	434	96.54	96.54	96.54	96.77
PTKQU	86.15	61	75.41	78.69	78.69	78.69
PTKFILL	88.89	135	89.63	90.37	89.63	91.11
PTKINI	95.87	44	70.45	70.45	70.45	70.45
PTKONO	90.16	29	0.00	0.00	0.00	0.00
PTKPH	61.54	10	0.00	0.00	0.00	0.00
PTKRZ	45.45	38	78.95	78.95	78.95	78.95
TOTAL	88.22	751	86.15	86.55	86.42	86.82
<i>normalised</i>						
POS	IAA (%)	freq.	base	LDA	HGC	LDA+HGC
PTKANT	89.59	436	95.18	95.18	95.41	95.87
PTKQU	86.15	61	93.44	93.44	93.44	93.44
PTKFILL	88.89	135	94.81	96.30	96.30	96.30
PTKINI	95.87	44	77.27	77.27	77.27	77.27
PTKONO	90.16	29	0.00	0.00	0.00	0.00
PTKPH	61.54	10	30.00	30.00	30.00	30.00
PTKRZ	45.45	38	89.47	89.47	92.11	92.11
TOTAL	88.22	751	89.11	89.38	89.64	89.91

Table 12: Tagging results for the discourse markers (please note that the IAA was not computed on the same data but on a larger test set for the same POS tags and is thus not directly comparable).

For the two most frequent error tags, nouns (NN) and adjectives (ADJD), the combination of the LDA clustering and the HGC dictionary approach further reduce the error rate, showing that the two methods can provide complementary information. The same is true for *v*VFIN, PIS and PTKVZ.

Figure 3 shows the same analysis for the normalised data. Here, the most frequent errors made by the base tagger involved adjectives (ADJD), adverbs (ADV), proper names (NE), foreign material (FM), unfinished words (XYB), indefinite substitutive pronouns (PIS), finite verbs (*v*VFIN), past participles (*v*VPP), verb particles (PTKVZ), substitutive relative pronouns (PRELS), imperatives (*v*VIMP), demonstrative pronouns (PDS), determiners (ART), prepositions (APPR) and onomatopoeia and forms of echoism (PTKONO).

On the normalised data, the LDA model again results in a higher error rate for determiners and also causes a higher number of errors on demonstrative pronouns. The HGC model performs worse on foreign material, unfinished words, finite verbs and verb particles. Not surprisingly, the domain adaptation approaches have a higher impact on the original transcripts than on the normalised data.

Table 12 shows results for the discourse marker tags. To give an idea how humans perform on the same tags, we added the scores for IAA from Section 3.2 but would like to remind the reader that these scores have been obtained on a different test set and are thus not directly comparable.

At first glance the tagger does better than our human annotators. This, however, is only true for those tags with a strong most-frequent-sense baseline where the tagger has a strong bias for assigning the most frequent tag. The annotation of *ja* (yes) is a case in point. There are 206 instances of *ja* in the development set. Out of those, 7 instances are used as an interjection, 8 instances as an utterance-initial discourse particle, 29 instances of modal particles positioned in the middle field, and 162 answer particles.

All instances of *ja* as an answer particle have been tagged correctly by the tagger. However, utterance-initial discourse particles are either assigned the ADV or the PTKANT tag, and only one instance of the *ja* interjections received the correct tag while the other 6 had been annotated as PTKANT. This most-frequent-sense strategy results in an overall accuracy for PTKANT, PTKFILL and PTKRZ which is higher than the one of the human annotators. However, it would be wrong to claim that the tagger has learned to distinguish between these tags.

The tags with the lowest frequency have been ignored completely by the tagger when tagging the transcripts (PTKONO, PTKPH). On the normalised data, the tagger does at least tag some instances of the placeholder particle correctly.

5.5 Simple, lexicon-based normalisation

As seen above, there is still a huge gap between the results on the transcripts and those on the normalised version of the data. While our base tagger showed a difference in accuracy of around 6% on the transcripts and on the normalised data, the combination of the word clustering approach and the HGC dictionary method reduced the gap to around 5%.

We now try to further improve results on the transcripts by applying a normalisation dictionary extracted from the KiDKo corpus. Our dictionary has 14,030 entries, sorted by frequency. Our approach is very simple. For each word in the transcripts, we extract its most frequent normalisation and replace the word by its normalised form. We also remove colons (used to indicate lengthened vowels) from the word forms not found in the normalisation dictionary. This very simple approach gives us a further improvement of around 1% on the development set, increasing accuracy for our best setting (LDA + HGC) from 90.93 to 91.94%.

5.6 Final results

Finally, we validate the results on the held-out test set. Table 13 shows results for the different settings on the development and the test set. For convenience, we also repeat the two baselines (TreeTagger) and the results from Section 4.2 for our base tagger (without domain adaptation).

Results on the test set are in the same range as the ones on the development set. Best results are obtained by the combination of the word clustering approach and the HGC dictionary method and, for the transcripts, by applying normalisation using the simple dictionary approach. The last row of Table 13 shows results for our proposal for a coarse-grained classification, subsuming answer particles, interjections, question

	trans.		norm.	
	dev	test	dev	test
Baseline 1 (TreeTagger, original)	42.54	42.48	74.53	73.67
Baseline 2 (TreeTagger, re-trained)	58.42	59.90	91.76	91.56
Base tagger (lowercase)	87.24	88.43	93.95	94.14
LDA 50-10	88.95	89.53	95.02	94.89
HGC	90.66	90.81	95.86	95.66
LDA 50-10 + HGC	90.93	91.09	95.97	95.77
LDA 50-10 + HGC + normalisation	91.94	91.97	-	-
LDA 50-10 + HGC + norm., coarse (DM)	92.33	92.37	96.20	95.95

Table 13: Baselines and results for the different approaches on the development and test set

particles, backchannel signals, fillers and utterance-initial discourse particles under the label DM (Table 4). The coarse-grained classification does not have a strong impact on the overall results but slightly increases the accuracies by about half a percent.

6 Conclusions and future work

In this paper, we presented an extension of the STTS for the annotation of spoken language. Our extended tagset captures silent and filled pauses, backchannel signals, question tags, utterance-initial discourse particles, non-inflected verb forms, placeholders for ambiguous material as well as tags for unfinished or uninterpretable words. We also added a new punctuation tag for abandoned or interrupted utterances.

We showed that the new tagset can be applied by human annotators without causing an overall decrease in accuracy. We identified and discussed problematic tags and proposed a two-way classification scheme which comprises a coarse-grained tag for discourse markers, thus allowing one to consistently annotate spoken language data without spending too much time on difficult pragmatic distinctions. The fine-grained classification, paying attention to the distinctions between different discourse markers, can be used, if need be, and the fine-grained tags can always be reduced to the umbrella tag DM for those who do not wish (or trust) the more detailed analysis. Our proposal also includes a restructuring of the STTS, renaming answer particles and interjections and grouping both in the *discourse marker* category.

On the basis of our manual annotations, we developed a CRF-based POS tagger for spoken language. We showed different methods to address the out-of-vocabulary problem of our data and presented tagging accuracies of close to 92% on the original transcripts and of close to 96% on the normalised version of the data.

Much more can be done to improve the tagger. A more sophisticated approach to normalisation, for instance, might take into account the immediate context of the word form, thus reducing noise introduced by faulty normalisations. The LDA word

clustering approach will most probably benefit from a larger amount of unlabelled data, and further experiments on the normalisation of the unlabelled data used as input for the word clustering might also improve results.

Acknowledgements

This work was supported by a grant from the German Research Foundation (DFG) awarded to SFB 632 “Information Structure” of Potsdam University and Humboldt-University Berlin, Project B6: “The Kiezdeutsch Corpus”. We gratefully acknowledge the work of our transcribers and annotators, Anne Junghans, Sophie Hamm, Jana Kiolbassa, Marlen Leisner, Charlotte Pauli, Nadja Reinhold and Emiel Visser. We would also like to thank the three anonymous reviewers for their insightful comments.

References

- Auer, P. and Günthner, S. (2005). Die Entstehung von Diskursmarkern im Deutschen – ein Fall von Grammatikalisierung? In Leuschner, T., Mortelmans, T., and de Groot, S., editors, *Grammatikalisierung im Deutschen (Linguistik - Impulse und Tendenzen; 9.)*, pages 335–362. Berlin: de Gruyter.
- Biber, D. and Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.
- Biemann, C. (2006). Unsupervised part-of-speech tagging employing efficient graph clustering. In *Proceedings of the 44th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics.
- Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER treebank. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories*, pages 24–42.
- Breiteneder, A., Pitzl, M.-L., Majewski, S., and Klimpfinger, T. (2006). VOICE recording – Methodological challenges in the compilation of a corpus of spoken ELF. *Nordic Journal of English Studies*, 5(2):161–188.
- Chrupała, G. (2011). Efficient induction of probabilistic word classes with LDA. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 363–372, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.
- Chrupała, G. (2012). Hierarchical clustering of word class distributions. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 100–104, Montréal, Canada. Association for Computational Linguistics.
- Crookes, G. (1990). The utterance and other basic units for second language discourse analysis. *Applied Linguistics*, 11:183–199.
- Dürscheid, C. (2003). Medienkommunikation im Kontinuum von Mündlichkeit und Schriftlichkeit. Theoretische und empirische Probleme. *Zeitschrift für angewandte Linguistik*, 38:37–56.

- Eynde, F. V., Zavrel, J., and Daelemans, W. (2000). Part of speech tagging and lemmatisation for the Spoken Dutch Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*.
- Fernandez, R. and Ginzburg, J. (2002). Non-sentential utterances in dialogue: A corpus-based study. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*.
- Fitschen, A. (2004). *Ein computerlinguistisches Lexikon als komplexes System*. PhD thesis, Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart.
- Foster, P., Tonkyn, A., and Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3):354–375.
- Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1 of *ICASSP*, pages 517–520, San Francisco, California, USA.
- Hirschmann, H., Doolittle, S., and Lüdeling, A. (2007). Syntactic annotation of non-canonical linguistic structures. In *Proceedings of Corpus Linguistics 2007*, Birmingham, UK.
- Koch, P. and Oesterreicher, W. (1985). Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36:15–43.
- Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Meer, D. (2009). "Unschärfe Ränder". Einige kategoriale Überlegungen zu Konstruktionen mit dem Diskursmarker *ja* in konfrontativen Talkshowpassagen. In Günthner, S. and Bücker, J., editors, *Grammatik im Gespräch*, pages 87–114. Walter de Gruyter, Berlin, New York.
- Okazaki, N. (2007). CRFsuite: a fast implementation of Conditional Random Fields (CRFs).
- Oostdijk, N. (2000). The spoken Dutch corpus. Overview and first evaluation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*.
- Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., and Schneider, N. (2012). Part-of-speech tagging for Twitter: Word clusters and other advances. Technical Report CMU-ML-12-107, Carnegie Mellon University.
- Rehbein, I., Hirschmann, H., Lüdeling, A., and Reznicek, M. (2012). Better tags give better trees - or do they? In *Proceedings of Treebanks and Linguistic Theory (TLT-10)*.
- Reznicek, M., Walter, M., Schmidt, K., Lüdeling, A., Hirschmann, H., Krummes, C., and Andreas, T. (2010). *Das Falko-Handbuch: Korpusaufbau und Annotationen*. Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin, Berlin.
- Sampson, G. (2000). *English for the computer: the SUSANNE corpus and analytic scheme*. Clarendon Press, Oxford.
- Schiller, A., Teufel, S., and Thielen, C. (1995). Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart, Universität Tübingen.

- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.
- Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the EACL SIGDAT-Workshop: From Text to Tags*.
- Schmid, H. and Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics, COLING '08*.
- Schwitalla, J. (1976). Dialogsteuerung. Vorschläge zur Untersuchung. In Berens, F. J., Jäger, K.-H., Schank, G., and Schwitalla, J., editors, *Projekt Dialogstrukturen. Ein Arbeitsbericht*, pages 73–104. Hueber, München.
- Schwitalla, J. (2006). *Gesprochenes Deutsch. Eine Einführung*. Erich Schmidt Verlag, Berlin.
- Selting, M., Auer, P., Barden, B., Bergmann, J., Couper-Kuhlen, E., Günthner, S., Quasthoff, U., Meier, C., Schlobinski, P., and Uhmannel, S. (1998). Gesprächsanalytisches Transkriptionssystem (GAT). *Linguistische Berichte*, 173:91–122.
- Skut, W., Krenn, B., Brants, T., and Uszkoreit, H. (1997). An annotation scheme for free word order languages. In *Proceedings of the 5th Conference on Applied Natural Language Processing, ANLP '97*, pages 88–95, Washington D.C., USA.
- Søgaard, A. (2010). Simple semi-supervised training of part-of-speech taggers. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics.
- Stegmann, R., Telljohann, H., and Hinrichs, E. W. (2000). Stylebook for the German treebank in VERBMOBIL. Technical Report 239, Seminar für Sprachwissenschaft, Universität Tübingen.
- Telljohann, H., Hinrichs, E., and Kübler, S. (2004). The Tüba-D/Z Treebank: Annotating German with a context-free backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 2229–2235.
- Teuber, O. (1998). fasel beschreib erwähn – Der Inflektiv als Wortform des Deutschen. *Germanistische Linguistik*, 26(6):141–142.
- Wiese, H., Freywald, U., Schalowski, S., and Mayr, K. (2012). Das KiezDeutsch-Korpus. Spontansprachliche Daten Jugendlicher aus urbanen Wohngebieten. *Deutsche Sprache*, 2(40):797–123.

7 Appendix

POS	TiGer	KiDKo	POS	TiGer	KiDKo
\$({	1159	226	PRF	172	77
\$.	1424	4744	PROAV	160	98
\$#	n.a.	846	PTKINI	n.a.	63
\$,	1539	891	PTKA	15	20
ADJA	1760	222	PTKANT	1	648
ADJD	566	887	PTKNEG	169	352
ADV	1187	2498	PTKONO	n.a.	39
APPO	12	2	PTKFILL	n.a.	212
APPR	2410	690	PTKPH	n.a.	12
APPRART	465	79	PTKQU	n.a.	94
APZR	13	0	PTKRZ	n.a.	62
ART	3124	627	PTKVZ	159	258
CARD	580	137	PTKZU	158	25
FM	27	104	PWAT	4	6
ITJ	0	592	PWAV	54	207
KOKOM	71	44	PWS	15	202
KON	713	596	TRUNC	36	5
KOUI	33	3	VAFIN	819	1368
KOUS	244	224	VAIMP	0	4
NE	1821	478	VAINF	121	35
NN	5856	1540	VAPP	51	6
PAUSE	n.a.	2475	VMFIN	258	334
PDAT	105	55	VMIMP	0	1
PDS	111	537	VMINF	18	3
PIAT	172	128	VVFIN	1090	1032
PIS	175	321	VVIMP	11	351
PTKINI	n.a.	1	VVINFL	424	410
PPER	454	2292	VVIZU	53	6
PPOSAT	188	196	VVPP	586	536
PPOSS	2	15	XY	24	0
PRELAT	13	2	XYB	n.a.	115
PRELS	205	47	XYU	n.a.	733

Table 14: Distribution of POS tags in a subset of TiGer (28,827 tokens) and in KiDKo (normalised training/development/test sets; 28,827 tokens)