Extending the STTS for the Annotation of Spoken Language

Ines Rehbein
SFB 632 "Information Structure"
German Departement
Potsdam University

Sören Schalowski
SFB 632 "Information Structure"
German Departement
Potsdam University

{irehbein,soeren.schalowski}@uni-potsdam.de

Abstract

This paper presents an extension to the Stuttgart-Tübingen TagSet, the standard part-of-speech tag set for German, for the annotation of spoken language. The additional tags deal with hesitations, backchannel signals, interruptions, onomatopoeia and uninterpretable material. They allow one to capture phenomena specific to spoken language while, at the same time, preserving inter-operability with already existing corpora of written language.

1 Introduction

Language resources annotated with part-of-speech (POS) information are a valuable resource for linguistic studies as well as for research in the humanities in general. Most existing corpora for German, however, include only written language data, often from the domain of newspaper text.

Recent years have seen an increasing interest in building language resources with data from a variety of domains like spoken language, historical language or computer-mediated communication. This has started a discussion on best practices for annotating and processing *non-canonical* language,[1] where *non-canonical* refers to all kinds of language data which deviate from standard written text. Important issues which have been addressed are the need for normalisation a) to enable corpus searches for all (pronounciation or spelling) variants of one word token, and b) to support the use of off-the-shelf NLP tools developed for written text. Another topic of discussion is the adequacy of existing annotation schemes for new types of data (e.g. a new POS tag set for annotating Twitter data (Gimpel et al., 2011)).

The main objective for using existing annotation schemes for annotating a new variety of data is *inter-operability* with existing resources. There are two aspects of inter-operability. First, we want to be able to use different corpora in cross-linguistic studies and compare results obtained from different corpora, which is only possible if all resources employ the same annotation scheme. Second, we would also like to use existing off-the-shelf NLP tools for the semi-automatic annotation of new data, which again would not be possible when using newly developed annotation schemes for which no training data is available.

We acknowledge the importance of the first objective while, at the same time, arguing for the need to provide a more adequate linguistic description of spoken language phenomena on the POS level. Areas of application are linguistic investigations of e.g. communication strategies or disfluencies in language production, amongst others. We thus propose an extension to an existing tag set with additional tags for phenomena not yet covered by the annotation scheme. This approach guarantees the comparability with other corpora using the original tag set while providing the means for a more adequate description of spoken language.

[1] See, e.g., the workshop on *Annotation of Corpora for Research in the Humanities* (ACRH), the LREC 2012 workshops *Best Practices for Speech Corpora in Linguistic Research, Adaptation of Language Resources and Tools for Processing Cultural Heritage Objects, NLP can u tag #user_generated content?!*, or the NAACL 2012 workshop on *Syntactic Analysis of Non-canonical Language*.

2 Related Work

Previous work on POS tagging spoken language mostly relies on existing annotation schemes developed for written language, using only minor additions (if any). The Switchboard corpus (Godfrey et al., 1992) provides a fine-grained annotation of disfluencies in spoken dialogues. On the POS level, however, the annotations do not distinguish between interjections, backchannel signals, answer particles, filled pauses or other types of discourse particles. The same is true for the Corpus of Spoken Netherlands (CGN) (Schuurman et al., 2003) and the spoken part of the BNC (Burnard, 2007). Nivre and Grönqvist (2001) extend a tagset developed for written Swedish with two tags designed for spoken language (*feedback* for answer particles and adverbs with similar function, and *own communication management* for filled pauses).

The only linguistically annotated, publicly available corpus of spoken German we are aware of is the Tübingen Treebank of Spoken German (TüBa-D/S) (Stegmann et al., 2000). The TüBa-D/S was created in the Verbmobil project (Wahlster, 2000) and is annotated with POS tags and syntactic information (phrase structure trees, grammatical dependencies and topological fields (Höhle, 1998)).

2.1 POS annotation in the TüBa-D/S

The TüBa-D/S uses the Stuttgart-Tübingen TagSet (STTS) (Schiller et al., 1995), the standard POS tag set for German which was also used (with minor variations) in the creation of the three German newspaper treebanks, NEGRA (Skut et al., 1998), TIGER (Brants et al., 2002) and TüBa-D/Z (Telljohann et al., 2004).

There are a number of phenomena specific to spoken language which are not captured by the STTS, including hesitations, backchannel signals, question tags, onomatopoeia, and non-words. As the TüBa-D/S does not use additional POS tags to label these phenomena,[2] it is interesting to see how they have been treated in the corpus.

Concerning hesitations, the TüBa-D/S encodes neither silent nor filled pauses such as *ahm, äh* (uhm, er). Occurences of these seem to have been removed from the corpus. Particles expressing surprise (*ah, oh*), affirmation such as *gell* (right), or discourse particles such as *tja* (well) have been included in the transcripton and assigned the label for interjections (ITJ). Backchannel signals as in (1) are also annotated as interjections in TüBa-D/S .

(1) A: also ab zwölf Uhr habe ich bereits
well from twelve o'clock have I already
einen Termin
a date

B: **mhm** welche Uhrzeit
mhm what time

A: Well, from 12:00 on I already have a date.
B: Mhm, what time?

Question tags like *nicht/ne* (no), *richtig/gell* (right), *okay* (okay), *oder* (or) have been labelled as interjections, too (Example 2).

(2) A: es war doch Donnerstag , **ne** ?
it was however Thursday , no ?
It was Thursday, right?

As a result, there is no straight-forward way to search for occurrences of these phenomena in the corpus. This is due to the fact that the Verbmobil corpus was created with an eye on applications for machine translation of spontaneous dialogues, and thus phenomena specific to spoken language were not the focus of the annotation.

3 Extending the STTS for spoken language

Our extension to the STTS provides 11 additional tags for annotating spoken language phenomena (Table 1).

3.1 Hesitations

Our extended tag set allows one to encode silent pauses as well as filled pauses.

POS	description	POS	description
PTKFILL	*particle, filler*	PAUSE	*pause, silent*
PTK	*particle, unspec.*	NINFL	*inflective*
PTKREZ	*backchannel*	XYB	*unfinished word*
PTKONO	*onomatopoeia*	XYU	*uninterpretable*
PTKQU	*question tag*	$#	*unfinished*
PTKPH	*placeholder*		*utterance*

Table 1: Additional POS tags for spoken language data

[2] The only additional tag used in TüBa-D/S is the BS tag used for isolated letters, which is not defined in the STTS.

The **PAUSE** tag is used for silent (unfilled) pauses which can occur at any position in the utterance.

(3) das ist irgend so ein **(-)** Rapper
this is some so a rapper
This is some eh rapper.

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

(4) das ist irgend so ein **äh** Rapper
this is some so a eh rapper
This is some rapper.

3.2 Other particles

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

(5) das Lied ging so **lalalala**
the song went like lalalala

The **PTKREZ** 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.

(6) A: stell dir das mal vor !
A: imagine you this PART. VERB PART. !
Imagine that !

(7) 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* (yes). To support consistency of annotation, we always label *ja* as an answer particle and not as a backchannel signal.

The **PTKQU** tag is used for question tags like *nicht/ne* (no), *richtig/gell* (right), *oder* (or), added to the end of a positive or negative statement.

(8) wir treffen uns am Kino , **ne** ?
we meet REFL at the cinema , no ?
We'll meet at the cinema. Right ?

The **PTK** tag is used for unspecific particles such as *ja* (yes), *na* (there, well) when occuring in utterance initial position.

(9) **ja** wer bist du denn ?
yes who are you then ?
And who are you now?

Please note that most occurences of *ja* (yes) in the middle field are modal particles (Example 10) which are assigned the ADV label (adverb) in the German treebanks. Occurences of *ja* in the prefield, on the other hand, should be considered as discourse markers and thus should be treated differently (also see Meer (2007) for a discussion on the different word classes of *ja*).

(10) die hat **ja** auch nicht funktioniert .
this has PTK.MOD also not worked .
This didn't work, either.

The **PTKPH** tag is used as a placeholder when the correct word class can not be inferred from the context. Example (11), for instance, has many 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).

(11) er hat **dings** hier .
he has thingy here .

a. er hat MP3-Player$_{NN}$ hier .
he has MP3 player here .

b. er hat gewonnen$_{VVPP}$ hier .
he has won here .

c. er hat (Schuhe gekauft)$_{VP}$ hier .
he has shoes bought here .

3.3 Non-words

Our tag set distinguishes 3 types of non-words.

1. uninterpretable
2. non-word in abandoned utterances
3. other

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. This tag should also be used for word tokens where it is not clear whether they are unfinished or simply non-words.

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

The **XYB** tag is used for abandoned words.

(13) ich **ha** # sie kommt **Sams** äh Sonntag .
I ha- she comes Satur- eh Sunday .
I ha- she'll come on Satur- eh Sunday.

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

3.4 Inflective

The **NINFL** tag is used for non-inflected verb forms (Teuber, 1998) which are a common stilistic device in comics and computer-mediated communication, but are also used in spoken language.

(14) ich muss noch putzen . **seufz** !
I must still clean . sigh !
I still have to clean. Sigh!

3.5 Punctuation

The **$#** tag is used to mark interrupted/abandoned utterances. These can (but not necessarily do) include unfinished words, as in Example (15).

(15) sie war ge #
she was (UNINTERPRETABLE) #

4 Inter-Annotator Agreement

We measured inter-annotator agreement for three human annotators using the extended tagset on a test set (3415 tokens) of spontaneous multi-party dialogues from the KiDKo corpus (Wiese et al., 2012) and achieved a Fleiss' κ of 0.975 (% agr. 96.5). Many of the errors made by the annotators concern the different functions of *ja* in spoken data (discourse marker vs. answer particles).

5 Discussion

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 on second glance does it become obvious that what looked similar at first does not necessarily need to be the same.

A case in point is *ja* (yes), which in written text mostly occurs as a modal particle in the middle field, labelled as ADV, while in spoken dialogues occurences of *ja* in utterance-initial position, labelled as answer particles (PTKANT), are by far the more frequent (Table 2). Motivated by the difference in distribution, we took a closer look at these instances and observed that many of them

POS	TIGER	TüBa-D/Z	TüBa-D/S
PTKANT	43	147	27986
ADV	154	372	4679
ITJ	2	0	0
NN	0	16	0
total	199	536	32664

Table 2: Distribution of *ja* (yes) in different corpora, normalised by corpus size

are in fact discourse markers (Example 9). We thus added the label PTK to our tag set, which is defined by its position in the utterance and its function.

As a second example, consider *weil* (because, since) which, according to standard grammars, is a subordinating conjunction. In TIGER as well as in the TüBa-D/S, all occurences of *weil* are annotated as KOUS (subordinating conjunction). However, in TüBa-D/S we also find examples where *weil* is used to coordinate two main clauses (indicated by V2 word order) and thus should be labelled as a coordination (KON) (Example 16).

(16) [...] fahren wir nicht zu früh los , **weil** sonst
[...] drive we not too early PTK , because else
bin ich unausgeschlafen
am I sleepdeprived
let's not start too early, else I'll be tired out

Finally, it is important to keep in mind that all types of linguistic annotation 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 (see Example 11). 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.

6 Conclusion

In the paper we extended the Stuttgart-Tübingen TagSet, the standard POS tag set for German, for the annotation of spoken language. Our extension allows for a more meaningful treatment of spoken language phenomena while also maintaining the comparability with corpora of written text annotated with the original version of the STTS.

Acknowledgments

This work was supported by a grant from the German Research Association (DFG) awarded to the Collaborative Research Centre (SFB) 632 "Information Structure". We gratefully acknowledge the work of our annotators, Nadja Reinhold and Emiel Visser.

References

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002)*, Sofia, Bulgaria.

Lou Burnard. 2007. Reference guide for the British National Corpus XML edition. Technical report, http://www.natcorp.ox.ac.uk/XMLedition/URG/.

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

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 517–520, San Francisco, California, USA.

Hagen Hirschmann, Seanna Doolittle, and Anke Lüdeling. 2007. Syntactic annotation of non-canonical linguistic structures. In *Proceedings of Corpus Linguistics 2007*, Birmingham, UK.

Tilman Höhle. 1998. Der Begriff "Mittelfeld", Anmerkungen über die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses 1985*, pages 329–340, Göttingen, Germany.

Dorothee Meer. 2007. "ja er redet nur Müll hier." – Funktionen von 'ja' als Diskursmarker in täglichen Talkshows. *gidi Arbeitspapierreihe*, 11.

Joakim Nivre and Leif Grönqvist. 2001. Tagging a Corpus of Spoken Swedish. *International Journal of Corpus Linguistics*, 6(1):47–78.

Marc Reznicek, Maik Walter, Karin Schmidt, Anke Lüdeling, Hagen Hirschmann, Cedric Krummes, and Torsten Andreas, 2010. *Das Falko-Handbuch: Korpusaufbau und Annotationen*. Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin, Berlin.

Anne Schiller, Simone Teufel, and Christine Thielen. 1995. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart, Universität Tübingen.

Ineke Schuurman, Machteld Schouppe, Heleen Hoekstra, and Ton van der Woulden. 2003. CGN, an annotated corpus of spoken Dutch. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*.

Wojciech Skut, Thorsten Brants, Brigitte Krenn, and Hans Uszkoreit. 1998. A linguistically interpreted corpus of German newspaper text. In *Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation*, pages 705–711.

Rosmary Stegmann, Heike Telljohann, and Erhard W. Hinrichs. 2000. Stylebook for the German Treebank in VERBMOBIL. Technical Report 239, Seminar für Sprachwissenschaft, Universität Tübingen.

Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. 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.

Oliver Teuber. 1998. fasel beschreib erwähn – Der Inflektiv als Wortform des Deutschen. *Germanistische Linguistik*, 26(6):141–142.

Wolfgang Wahlster, editor. 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Berlin.

Heike Wiese, Ulrike Freywald, Sören Schalowski, and Katharina Mayr. 2012. Das Kiezdeutsch-Korpus. spontansprachliche Daten Jugendlicher aus urbanen Wohngebieten. *Deutsche Sprache*, 40(2):97–123.