STTS 2.0? Improving the Tagset for the Part-of-Speech-Tagging of German Spoken Data

Swantje Westpfahl
Institut für Deutsche Sprache, Mannheim
westpfahl@ids-mannheim.de

Abstract

Part-of-speech tagging (POS-tagging) of spoken data requires different means of annotation than POS-tagging of written and edited texts. In order to capture the features of German spoken language, a distinct tagset is needed to respond to the kinds of elements which only occur in speech. In order to create such a coherent tagset the most prominent phenomena of spoken language need to be analyzed, especially with respect to how they differ from written language. First evaluations have shown that the most prominent cause (over 50%) of errors in the existing automatized POS-tagging of transcripts of spoken German with the Stuttgart Tübingen Tagset (STTS) and the treetagger was the inaccurate interpretation of speech particles. One reason for this is that this class of words is virtually absent from the current STTS. This paper proposes a recategorization of the STTS in the field of speech particles based on distributional factors rather than semantics. The ultimate aim is to create a comprehensive reference corpus of spoken German data for the global research community. It is imperative that all phenomena are reliably recorded in future part-of-speech tag labels.

1 Introduction

In the Institute for German Language (Institut für Deutsche Sprache, IDS Mannheim) a large reference corpus of German spoken data is currently being built. It already contains more than 100 hours of transcribed audio material, i.e. about one million tokens. The aim of my dissertation is to annotate the corpus with Part-of-Speech-tags (POS-tags) and thus to tackle the theoretical problems which originate from the differences between spoken and written language. On the one hand, as the corpus is growing fast, ways must be found to automate this, i.e. without manual correction. On the other hand there are no tools to accomplish such a task at present. First tests running the treetagger (Schmid 1995) with the Stuttgart Tübingen Tagset (STTS) (Schiller et al. 1999), which on written data show an accuracy of 97.53% (Schmid 1995) have shown that the accuracy of these tools on spoken data is far below acceptable, i.e. they only show an accuracy of 81.16% (Westpfahl and Schmidt 2013). There are two main reasons for this. First of all, the structure of German spoken language is quite different from the structure of German written language due to many elliptic structures, disruptions, repetitions etc. Furthermore, punctuation is not annotated in the corpus; hence the algorithms have no "proper sentences" to work with.

As was shown in the studies of Westpfahl and Schmidt (2013) and Rehbein and Schalowski (2013), the mistakes in annotating POS-tags on German spoken language are due to a lack of suitable categories in the tagset; namely categories which reflect the manifold speech particles, vernacular use of pronouns, verbs and items which are impossible to categorize grammatically. As can be seen in Table 1, a first analysis of tagging errors showed that more than 50% of the mistakes are due to mis-tagged discourse markers, interjections and speech particles. Hence, to reach the goal of an automatized POS-tagging, the tagset must firstly be adapted to those phenomena. It is the aim of this paper to provide a theoretical foundation on how to comply with the need to re-categorize the existing tag set for speech particles; creation of new tags and merging of existing tags are both proposed as seems relevant for the particular data. Problems which are due to verbs, pronouns and non-words cannot be discussed here.

Table 1: Errors in POS annotation > 5% (Westpfahl and Schmidt 2013)

Errors in POS annotation > 5%	
Particles and interjections	51,59%
Pronouns	13,43%
Verbs	9,14%
XY non words	8,18%

LAW VIII - The 8th Linguistic Annotation Workshop, pages 1–10,
Dublin, Ireland, August 23-24 2014.

2 Related work

Various spoken language corpora which are annotated with POS tags already exist. English language corpora are for example the BNC (Burnard 2007), the Switchboard corpus (Godfrey et al. 1992), Vienna-Oxford International Corpus of English (VOICE) (VOICE 2013) and the Christine Corpus (Sampson 2000). While both the Switchboard corpus and the BNC use POS tag-sets developed for written data, VOICE and the Christine corpus adapted theirs specifically for spoken language. VOICE added 26 POS tags to the Penn treebank POS tagset which include tags for, among others, discourse markers or response particles and also non-verbal elements like laughter and breathing. The Christine corpus uses a very fine grained tagset with more than 400 tags annotating morpho-syntactic as well as rich pragmatic information.

A POS tagset which has been especially designed for a corpus of spoken language is the tagset of the Spoken Dutch Corpus (Oosterdijk 2000). However, although the tagset consists of more than 300 tags, all discourse related items are tagged as interjections.

For the German language there is the Tübingen Treebank of Spoken German (TüBa-D/S), which uses the STTS with no alterations whatsoever (Telljohann et al. 2012).

In order to find a solution on how to tag non-standard texts with the STTS, an interest group was set up in 2012. Within this interest group a work group formed which especially focused their attention on the adaption of the STTS for spoken language and computer mediated language (CMC), namely for the corpora Kietz-Deutsch Corpus, the Dortmund chat corpus and our corpus (DGD2/FOLK). As a first result, three papers were published with some suggestions on which phenomena should be represented in an adapted tagset (Rehbein and Schalowski 2013; Bartz et al. 2013; Westpfahl and Schmidt 2013). The present paper is meant to give an overview of a theoretical foundation on how to comply with the need to recategorize the tagset with respect to speech particles, as so far only a “purely data-driven” approach has been discussed (Rehbein and Schalowski 2013).

3 Speech particles in the original STTS

The Stuttgart Tübingen Tagset (STTS) was conceptually developed for a corpus of newspaper articles and only those classes of words which were frequently used were represented in the tagset. Therefore, modal particles, speech particles or discourse markers were not at the center of attention of Schiller et al. (1999) as their use in written texts is commonly understood to be 'bad style'. To understand the changes I have made in the tagset I shall first present the categories used for particles and discourse markers in the original tagset:

Table 2 categories for speech particles in the original STTS

Tags	Description	Example	Literal English translation
PTKVZ	verbal particles	[er gab] auf	[he gave] up
PTKZU	particle used with infinitives	zu [gehen]	to [walk]
PTKA	particle used with adjectives or adverbs	am [schönsten], zu [schnell]	most [beautiful], too [fast]
PTKNEG	negation particle	nicht	not
PTKANT	response particles	ja, nein, danke, bitte	yes, no, thanks, please
ITJ	interjections	mhm, ach, tja	uhum, oh, well

As one can see, the STTS is structured hierarchically; for particles, the basis tag would be "PTK" for "Partikel" and there are five subcategory tags. Furthermore, the tagset provided one category for interjections: ITJ, defining them after Bußmann (1990) as words which serve to express emotions, swearing and curses and for getting in contact with others. Formally they are invariable and syntactically independent from the sentence as well as having, strictly speaking, no lexical meaning (Schiller et al. 1999, S. 73).

Concerning modal particles, intensity particles or focus particles etc., the guidelines published with the tagset do not assign them their own category. It is implicitly clear from the cited examples that modal particles or intensity particles are to be tagged as adverbs "ADV". (Schiller et al. 1999)

On running the STTS with the treetagger on three transcripts of German spoken data (11029 tokens), one finds that 35.76% of all corrected items were incorrectly tagged as adverbs and yet again 85.87% of those items tagged as adverbs were actually particles or interjections (Westpfahl and Schmidt 2013). Thus, the first step in restructuring the tagset would be finding categories differentiating adverbs from particles as well as interjections and discourse markers.

4 Features of spoken German - Speech particles in German grammar references

In order to explain the categorization employed in our proposed STTS 2.0 one has to take a deeper look at how transcripts of spoken German differ from 'normal' written language. First of all, in our corpus, no punctuation is annotated and there also is no annotation on where a speaker's turn starts or ends. Secondly, it is typical of spoken language that not all utterances form "proper" sentences but are quite often disrupted, e.g. marked by extensions or anacolutha, apokoinu-constructions, repairs etc. All this would be represented as such in the transcripts.

Furthermore, there are also differences in the choice of words, i.e. some closed categories contain other or more tokens in spoken language and some speech phenomena simply do not occur in written language except for, maybe, in quoting direct speech. Some of those phenomena are even hard to describe as syntactic categories, e.g. hesitation markers or backchannel signals. Nevertheless, exactly those phenomena are particularly interesting in working with a corpus of spoken language.

The approach used for finding categories was to first take a look at the canon of German grammars and then check whether the classifications made there could be applied for the corpus data.

The most consulted grammars for the German language are (Duden 2006), (Zifonun 1997), (Engel 2004), (Helbig 2011), (Hoffmann 2013), (Weinrich 2005) and the online grammar grammis 2.0 (Institut für deutsche Sprache 2013). The most consulted articles dealing with speech particles are (Burkhardt 1982), (Hentschel and Weydt 2002), (Schwitalla 2012) and (Diewald 2006).

Looking at this literature it becomes obvious that research on this topic has, so far, not lead to a unified classification of speech particles, but rather to a plethora of classifications and concepts differing at times quite radically in definition and nomenclature. Even the terminology and definitions used for the supercategory 'particles' vary quite drastically. For some, particles are all word classes which do not inflect, hence conjunctions and prepositions would be counted as particles as well (Engel 2004). Others differentiate between 'particles sensu lato' (particles in the wider sense) and 'particles in the strict sense' or synsemantica (Hentschel and Weydt 2002; Duden 2006; Burkhardt 1982). Yet again others differentiate between those which distributionally contribute to the compositional structure of the sentence and those which can form sentence-independent units (Diewald 2006; Weinrich 2005; Hoffmann 2013; Zifonun 1997; Institut für deutsche Sprache 2013). For those 'sentence-independent' units, e.g. interjections and response particles, yet again a variety of terms is used: interactive units ("Interaktive Einheiten") (Hoffmann 2013; Zifonun 1997; Institut für deutsche Sprache 2013), discourse particles (Diewald 2006), speaker signals and particles of the dialogue ("Sprechersignale und Dialogpartikeln") (Weinrich 2005) or 'words of speech' ("Gesprächswörter") (Burkhardt 1982).

As for statistical POS tagging the most important feature is distribution, the differentiation between sentence-independent particles and sentence-internal particles seems to be a reasonable basis for classification. Hence we propose these two major categories for the tagset. However, there are also particles which are neither sentence-independent nor sentence-internal but are either in the pre-front field or in the end field of a sentence, namely discourse particles ("Diskurspartikeln"). Quite surprisingly, these phenomena are hardly mentioned in any standard grammar reference at all. The DUDEN (2006), Weinrich (2005), Burkhardt (1982) and Diewald (2006) subsume e.g. reinsurance signals ("Rückversicherungssignale") and starting signals ("Startsignale") under the term structuring particles ("Gliederungspartikeln"), however, no distinction is made on whether they can stand independently from the sentence or not (Duden 2006). The other grammars simply do not mention them at all. Nevertheless, a differentiation can be made between sentence-independent, sentence-internal and sentence-external particles. By 'external', I mean that they are not part of the core sentence yet 'need' the sentence. So how can these categories be subclassified now and which phenomena fall into these classes?

4.1 Non-grammatical or sentence-independent elements

Regarding those particles which are sentence-independent, e.g. "ähm" or "hmm", it is crucial to bear in mind that these phenomena cannot be classified according to their distribution or any syntactic features. Hence, the only criterion by which to differentiate them is with respect to their pragmatic function. Taking a look at the grammar reference canon, one finds that DUDEN (2006) differentiates between interjections and onomatopoeia and subclassifies the former ones into simple and complex interjections, i.e. between those which have homonyms in other word classes and those which do not. The GDS (Zifonun 1997), Hoffmann's grammar (2013) and grammis 2.0 (Breindl and Donalies 2012)

differentiate between interjections and response particles (“Responsive”) and Engel’s grammar (2004) adds initiating particles (“Initiativpartikeln”) and reaction particles (“Reaktive Partikeln”) to those. In contrast to that, Harald Weinrich’s grammar (2005) only defines interjections, but subclassifies those into situational, expressive and imitative interjections. Just looking at the terminology used for their classifications, one can get a hint of how contradictory the various definitions of interjections are. Whether response particles, onomatopoeia, inflectives or filled pauses are all interjections or separate classes of their own always depends on how broad or strict a definition for the interjection would be.

4.2 Sentence-external elements

Sentence-external particles, namely discourse markers (“Diskursmarker”) and tag questions (“Rückversicherungspartikeln”), are not classes which are explicitly named as such in any grammar reference yet are controversially debated in the research field of conversation analysis. Only the DUDEN uses the term “Diskursmarker” but not in describing it as word class of its own, but only to differentiate subjunctions from the use of the same lexeme (e.g. *weil* or *obwohl*) with main clauses. Nevertheless, in the grammars we do find classes which could be subsumed under these concepts even though they are classified as, for example, structuring particles (“Gliederungspartikeln”) (Hentschel and Weydt 2002; Burkhardt 1982), dialogue particles (“Dialogpartikeln”) (Weinrich 2005) or “Sequenzpartikeln” (sequencing particles) (Hentschel and Weydt 2002). However, in all of these classes no differentiation is made between those which are really distributionally bound to the pre-front field or the end field and those which are sentence-independent. In the literature on discourse markers there is no agreement on what is to be subsumed under that term either. Traugott (1997) and Auer and Günthner (2005) define them as every utterance which has a peripheral syntactical position and a ‘metapragmatic function’. What seems clear is that these phenomena came into existence through grammaticalization or degrammaticalization (Gohl and Günthner 1999; Brinton 1996; Günthner 2005; Leuschner 2005; Auer and Günthner 2005), hence most of them are homophones of adverbs, conjunctions, subjunctions etc. Imo (2012) yet again clearly differentiates between discourse markers and tag questions as, according to him, they have a different function, namely only to demand attention or sequencing turns whilst discourse markers would project the continuation of a speaker’s turn (Imo 2012).

4.3 Sentence-internal elements

Analyzing those particle categories, the only ones which seem to be quite indisputable are verbal particles and the ones which are defined by their form, i.e. the particle “zu” used with the infinitive (PTKZU), “am” preceding an adjective (PTKA) and the negation particle “nicht” (PTKNEG), although the online grammar grammis 2.0 defines it to be a subclass of the focus particle (Breindl and Donalies 2011).

Table 3 Comparison of criteria for modal particles and Abtönungspartikeln in the literature

grammar / criteria	DUDEN	HSK		GDS		Diewald		Schwitalla		Grammis	Hoffmann		Weinrich		Engel		Burkhardt	
	MP+AP	MP	AP	MP	AP	MP	AP	MP	AP	MP+AP	MP	AP	MP	AP	MP	AP	MP	AP
express speaker attitude	+		+		+	+		+		+	+	+	+		+			+
changing the illocution			(+)	+		+		+		+						+		+
changing the proposition		+		+		+					+		+					
answer for yes/no questions		+		+			n/a		n/a		+		-	n/a	+	-	n/a	
has constituency value		-			-	-				-		(-)	-					
may be negated		-			-					-						-		
can appear in front field	-		-		-	-		-		-		-	-		+	-		
always unstressed	+		+		+			+		+								

AP Abtönungspartikeln
MP modal particles
\+ criterion is explicitly mentioned
(+) criterion is implicitly mentioned
– criterion is explicitly denied
(–) criterion is implicitly denied

By contrast, table 3 shows that there is much disagreement on how to define or differentiate "Abtönungspartikeln" (I'm not able to find a translation; literally translated it would be something like 'shading' or 'coloration' particles, A/N) from modal particles ("Modalpartikeln"), such as German *mal, halt, doch,* or *ja*. Furthermore, there are differences on whether to make a distinction between these two terms, whether to treat them as synonyms (Duden 2006; Breindl and Donalies 2011a) or having only one class of items at all (Schwitalla 2012; Diewald 2006; Weinrich 2005; Burkhardt 1982).

Looking at the table one can find that the core definitions of those types of particles are very similar to each other. Criteria used to describe both types of particles in nearly all definitions are:

- the expression of attitudes, expectations, assumptions, and appraisal of the speaker and the addressee
- the inability to appear in the front field
- they never form constituents of a sentence and thus cannot be moved at all
- apart from a few exceptions, they cannot be stressed.

Also quite problematic is the differentiation of what is termed focus particles ("Fokuspartikeln"), scalar particles ("Gradpartikeln") and intensifying particles ("Intensitätspartikeln") such as German nur,sogar, sehr etc. as can be seen in table 4.

Table 4 Comparison of criteria for focus particles, scalar particles and intensifying particles in the literature

grammar / properties	DUDEN			grammis 2.0			GDS			HSK			Engel		
	FP	SP	IP[1]	FP	SP	IP	FP[2]	SP	IP	FP	SP[3]	IP	FP	SP	IP
modify NPs	+	–		(+)	(+)	–	+		–	n/a	n/a		n/a	+	n/a
may modify AdjPs, VPs, and words of number	+	+		(+)	(+)	+	+		+	n/a	n/a		n/a	+	n/a
scaling function	+	+		+	+	n/a	+		n/a	+	n/a		n/a	+	n/a
focus item they precede	+	n/a		+	+	n/a	+		n/a	+	n/a		n/a	+	n/a
intensifying or weakening function	n/a	+		n/a	n/a	+	+		+	n/a	+		n/a	+	n/a
grading function	(+)[4]	+		+	+	n/a	+		n/a	n/a	+		n/a	+	n/a
may be stressed	n/a	+		n/a	n/a	n/a	n/a		n/a	n/a	n/a		n/a	n/a	n/a
may appear in the front field	n/a	(–)[5]		–	–	–	(–)[6]		n/a	n/a	n/a		n/a	–	n/a
no effect on proposition of the sentence if omitted	(–)	+		n/a	n/a	+	+		n/a	–	n/a		n/a	n/a	n/a

FP focus particles
SP scalar particles
IP intensifying particles
+/(+)/–/(–) see table 3

It becomes obvious that the definitions of this group of particles are quite similar to each other except for the focus particles not being omittable without changing the proposition of the sentence and intensifying particles not being able to precede noun phrases. However, looking at spoken language, the last definition can be easily contradicted – e.g. considering the statement: "Das ist aber *sehr* fünfzehntes Jahrhundert" (This is *very* fifteenth century). Looking at the data of our corpus, quite a lot of examples come up where one could not easily decide whether the particles used would be only used for intensifying, bringing something into focus or for their scaling function, e.g.: "weil ich bin jetzt *echt* müde" (FOLK_E_00002_SE_01_T_01_DF_01, 00:30:46.94 - 00:30:53.93) (because I'm really tired now) or "*voll* die sau" (FOLK_E_00021_SE_01_T_16_DF_01, 02:07:20.34 - 02:07:26.09) (truly/utterly a pig).

A third group of phenomena which are quite inconsistently classified are connective particles ("Konnektivpartikeln"), maneuvering particles ("Rangierpartikeln") and conjunctive adverbs ("Konjunktionaladverbien") such as German *allerdings, deshalb* etc. On grammis 2.0 for example it is stated that the term connective particles is simply a synonym of maneuvering particles and conjunctive adverbs (Breindl and Donalies 2010). These terms are also used in the DUDEN (2006) and in (Engel

1 DUDEN claims that intensifying particles are synonyms to scalar particles.
2 The GDS claims that focus particles are synonyms to scalar particles.
3 The HSK claims that intensifying particles are synonyms to scalar particles.
4 DUDEN vaguely claims that only 'some' items of the class have a grading function.
5 DUDEN vaguely claims that 'most' of them can stand in the pre-front field.
6 GDS claims that they cannot stand in the pre-front field except for the words "noch" and "schon".

2004), however, they define them as something different than what is defined as connective particles in Zifonun (1997), Hoffmann (2013), Breindl and Donalies (2010). The problem here is based on the linguistic level on which they are defined. Some grammars define them according to their semantics and classify them as to their conjunctive function albeit not being conjunctions. Others define them according to their distribution in which case they rather have to be classified as adverbs rather than particles.

5 The STTS 2.0 – new categories for speech particles

5.1 Preliminary considerations

Restructuring the tagset is a task which requires some thoughts in advance. First of all, as the tagset is structured in a hierarchical way, new categories must fit into that hierarchical system. Secondly, as the aim of the restructuring is not just a theoretical one but aims at practical use for the research community, it needs to be comprehensible. However, the aim is not to follow a single grammar or theory but rather to build an unambiguous system of categories which are mutually exclusive and allow for an exhaustive categorization when applied to data of transcribed spoken language. The general principles followed were to construct the tagset as detailed as possible, to allow the research community to find as many phenomena typical for spoken language as possible, yet as coarse as necessary in order to maintain consistency and to create mutually exclusive categories. In contrast to, say, the VOICE corpus, in which several possible word class categories can be assigned to a token if it is ambiguous (VOICE 2013), in an automatized tagging relying only on statistical values, ambiguity – especially with respect to pragmatic information – cannot be taken into account as the tagging will not be manually corrected. Consistent with the original guidelines of the STTS each item shall receive only one tag (Schiller et al. 1999). As a result, multiword expressions will not be tagged as one item either, even when the pragmatic information in such cases might be lost. However, the new structures built should be coherent concerning the linguistic levels on which the annotation is based. It has been discussed whether e.g. pragmatic and syntactic information should be specifically annotated on different annotation levels (see e.g. Rehbein 2013). One of the main reasons to annotate only one level of POS tags in our corpus is that, looking at spoken data, the syntactic function of an item is often deeply intertwined with its pragmatic function. Nevertheless, this paper suggests a reclassification which on a theoretical level aims at a clear representation of the distinction between the linguistic levels which shall be assigned through POS tags.

In addition, as the transcripts are based on spoken language and follow the cGAT conventions, one has to take into account that there are many utterances transcribed which cannot even be seen as 'words', like sighing, laughing or breathing. Hence, the categories created for typical spoken language phenomena will still adhere to the concept of a “word” and only those items shall receive a POS tag.

Being aware that the classical concept of the sentence cannot be applied to these transcripts of spoken data, the concept of the verbal bracket (Verbklammer) is still fundamental for the new categorization in order to describe the items in the utterances syntactically and also to determine whether they apply to a syntactical concept at all.

5.2 Extensions to the STTS

An overview on the structure of the categorization is given in table 5. Firstly, items like e.g. hesitation markers, interjections, onomatopoeia, inflectives or backchannel signals cannot be looked at on a syntactic linguistic level as they are not part of the syntax of a sentence. They shall be tagged as non-grammatical elements and thus receive the supercategory tag NG. As one category for all non-grammatical items would hardly be satisfactory to depict these various typical spoken language phenomena, one needs to consider a different linguistic level in order to further categorize them; namely their pragmatic function.

To ensure that the subcategories are mutually exclusive, a closer look into the corpus data was necessary to check whether one (and only one) pragmatic function could be assigned to items that are considered non-grammatical. Wherever this was not the case and one item could have several pragmatic functions, the items would have to be categorized into a 'broader' class of items. Finally, items like onomatopoeia, inflectives and hesitation markers only have this one pragmatic function and thus get their own POS tag categories NGONO (Onomatopoetika), NGAKW (Aktionswörter) and NGHES

(Hesitationspartikeln). However, response particles, backchannel signals and interjections do quite often take each other's functions, e.g. in the following example it is not clear whether "ach" is used as a response particle, a backchannel signal or an interjection.

LB °h isch ne GUte frage, ((schmatzt)) °hh des hat einfach mit der diagNOse zu tun.
(that's a good question ((smacking lips)) it's just about the diagnosis)
(0.22)
LB gucke ma uns NAchher mal an dann.
(we'll have a closer look at that later)
ML ACH so;
(ah)
LB ja?
(yes) (FOLK_E_00008_SE_01_T_01_DF_01, 00:15:21.66 - 00:15:28.84)

Hence, although on a theoretical level there might be differences between those classes, in analyzing spoken language these differentiations cannot be made in every case. Thus, there will only be one POS tag for those items in the STTS 2.0 which have the function of signaling response, backchanneling or interjections – the NGIRR for "Interjektionen, Rezeptionssignale und Responsive". Obviously, what formerly has been tagged as answering particle PTKANT (Antwortpartikel) will subsequently be tagged as NGIRR. This restructuring needs to be done as the response particles "yes", "no", "maybe" etc. are not – like the other particles which are tagged with the supercategory PTK – syntactically integrated in the sentence, i.e. located in the middle field of a sentence.

Secondly, there is the group of speech particles which are not part of the core sentence construction, yet pragmatically cannot stand on their own. These 'sentence external' (SE) elements can be subclassified into two classes. Discourse markers stand in the pre-front field and need a sentence to follow, i.e. they open up a projection which needs to be filled by the following. Tag questions stand in the end field and are used to raise the hearer's attention. Hence, two new POS tags are introduced to tag those items: SEDM (discourse particles) and SEQU (tag questions).

Table 5 schematic overview on the reclassification of speech particles

subject	POS tagging	distributional features	proposed tags	examples
Items in the corpus	**no tags assigned** No stable phonetical form, annotated according to cGAT conventions (e.g. sighing, laughing, breathing etc.).			((stöhnt)), ((lacht)), °hhh (sighs) (laughs) (breathing)
	tags assigned	**sentence-independent → non grammatical elements (NG)**	**NGIRR** interjections, response signals and backchannel behavior	ach, ja, hmhm (oh, yes, uhum)
			NGHES hesitation signals	äh, ähm (uhh, uhm)
			NGAKW action words (inflectives)	lol, grins, seufz (lol, grin, sighing)
			NGONO onomatopoeia	muh, miau, kikeriki (moo, meow, cock-a-doodle-doo)
		dependent on grammatical constructions yet not part of them → sentence-external elements (SE)	**SEDM** discourse particles	also [ich glaube …] (well [I think])
			SEQU tag questions	[ist gut] ne? ([it's good] isn't it?)
		sentence-internal → particles (PTK) (other than PTKZU, PTKA, PTKNEG, PTKVZ)	**PTKIFG** intensifying, focus, and scalar particles	sehr [schön], nur [sie], viel [mehr] (very [nice], only [her], much [more])
			PTKMA modal particles und Abtnungspartikeln	halt, mal, ja, schon (just, once)[7]
			PTKLEX particles which are part of a multi-word expression	[noch] eine/r, immer [noch] (another, still)

Finally, there are those speech particles which are syntactically integrated in the core sentence, i.e. are situated in the middle field. Those which are already represented in categories in the tagset and which are categorized based on their syntactic features will remain. The PTKANT tag will be removed from this category. Additionally, those sentence-internal particles which formerly have been tagged as

[7] There are hardly any Abtönungspartikeln in the English language, thus no literal translations are possible.

adverbs, i.e. modal, focus, scalar or intensifying particles shall be categorized as particles. Although the naming and the concepts for those particles are highly debated in the literature, syntactically, one can clearly differentiate them from adverbs as adverbs can stand in the front field on their own whilst speech particles cannot (Breindl and Donalies 2011). Moreover, as Hirschmann (2013) pointed out, one can divide all these speech particle concepts into two groups: those which can be moved to the front field together with their mother phrase and those which cannot be moved at all. The latter ones are either modal particles or Abtönungspartikeln, the former ones intensifying particles, focus particles and scalar particles (Breindl and Donalies 2011b). As evidently not even the grammars can give clear guidelines for the distinction of these classes, a categorization can only be based on distributional features. Consequently, there shall be two new POS tag categories PTKIFG (Intensitäts-, Fokus- und Gradpartikeln) and PTKMA (Modal- und Abtönungspartikeln). However, annotating data one comes across a set of sentence-internal particles which have not been accounted for so far. Hirschmann (2013) presented an analysis of items which are part of multi-word lexemes. They are bound to other lexemes not by modifying them as an intensifyer, focus, or scalar particle, but they have to be considered parts of multi-word constructions. This can be proven by the fact that the elements in question lose the meaning which they possess without the element they are joint with. From an orthographic point of view, however, those particles, together with the other item, build a phrasal constituent. For example, “immer” and “noch” in: “Baba ist *immer noch* brummelig” (Baba is still grumpy) (FOLK_E_00016, 13), together semantically form one lexeme which can also be seen in the translation where both together are translated as “still”. Crucially, the word “immer” is neither an adverb with the usual meaning “always” here, nor is it an intensifier with the meaning “increasingly” as in “immer besser” (increasingly well/better). In this one idiosyncratic case (“immer noch”), “immer” can onlybe interpreted together with “noch” which can only be moved as a multi-word lexeme in the sentence. The adverb “noch” can still be interpreted as the head of the whole expression. In this respect, lexicalized particles are similar to the group of PTKIFG, with the difference that they neither have an intensifying, scaling or focusing function. It seems like they are a very interesting group of particles, as one can analyze the gradual grammaticalization cline in such items. Finding that it is a very restricted group of possible items they shall receive their own tag PTKLEX for “particle in a multi-word lexeme”.

6 Conclusion and outlook

This paper presented a proposal as to how new tag categories for an improved version of the STTS (STTS 2.0) in the field of speech particles could look like.

To see whether these new categories work for part of speech annotation, guidelines have been written and the work on annotating a gold standard of about 100,000 tokens has begun. In order to evaluate and validate the proposed tagset and the guidelines, Cohen’s kappa will be used to assess the inter-annotator agreement. In addition, post-processing has been implemented that already helps to improve the accuracy of the output, e.g. by assigning POS-tags to those items which do not have any homonyms in other word classes, i.e. through a list of items which shall receive this tag. A first analysis shows that this proved to be extremely useful for the categories NGIRR, NGONO, NGHES, NGAKW and SEQU.

However, in order to fully automatize part of speech tagging of transcripts of spoken language, a retraining of the tagger will be necessary. Moreover – although the errors due to mis-tagged speech particles were the most prominent cause for the low precision rate – additional sources of errors will have to be analyzed to be able to create a coherent tagset for spoken language annotation. The analysis of the colloquial use of pronouns, verbs, foreign language material or in the STTS so called ‘non-words’ might call for a further recategorization of the Stuttgart Tübingen tagset.

References

Auer, Peter; Günthner, Susanne (2005): Die Entstehung von Diskursmarkern im Deutschen - ein Fall von Grammatikalisierung? In: Torsten Leuschner (Hg.): Grammatikalisierung im Deutschen. Berlin: De Gruyter (Linguistik - Impulse & Tendenzen), S. 335–362.

Bartz, Thomas; Beißwenger, Michael; Storrer, Angelika (2013): Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. In Journal for Language Technology and Computational Linguistics (28(1)), pp. 155–198.

Breindl, Eva; Donalies, Elke (2011): Abtönungspartikel. grammis 2.0. das grammatische informationssystem des instituts für deutsche sprache (ids). Institut für deutsche Sprache. Online verfügbar unter http://hypermedia.ids-mannheim.de/call/public/sysgram.ansicht?v_typ=d&v_id=392, zuletzt aktualisiert am 05.05.2011, zuletzt geprüft am 11.09.2013.

Breindl, Eva; Donalies, Elke (2011): Fokuspartikel. grammis 2.0. das grammatische informationssystem des instituts für deutsche sprache (ids). Institut für deutsche Sprache. Available online at http://hypermedia.ids-mannheim.de/call/public/sysgram.ansicht?v_typ=d&v_id=408, updated on 5/5/2011, checked on 8/20/2013.

Breindl, Eva; Donalies, Elke (2012): Intensitätspartikel. grammis 2.0. das grammatische informationssystem des instituts für deutsche sprache (ids). Institut für deutsche Sprache. Available online at http://hypermedia.ids-mannheim.de/call/public/sysgram.ansicht?v_typ=d&v_id=391, updated on 1/12/2012, checked on 8/20/2013.

Breindl, Eva; Donalies, Elke (2012): Interaktive Einheiten. grammis 2.0. das grammatische informationssystem des instituts für deutsche sprache (ids). Institut für deutsche Sprache. Online verfügbar unter http://hypermedia.ids-mannheim.de/call/public/sysgram.ansicht?v_typ=d&v_id=370, zuletzt aktualisiert am 12.01.2012, zuletzt geprüft am 13.01.2014.

Breindl, Eva; Donalies, Elke (2011): Konnektivpartikel. grammis 2.0. das grammatische informationssystem des instituts für deutsche sprache (ids). Institut für deutsche Sprache. Online verfügbar unter http://hypermedia.ids-mannheim.de/call/public/sysgram.ansicht?v_id=410, zuletzt aktualisiert am 05.05.2011, zuletzt geprüft am 08.10.2013.

Brinton, Laurel J. (1996): Pragmatic Markers in English: Grammaticalization and Discourse Functions. Berlin, Germany: Mouton de Gruyter (Topics in English Linguistics (TopicsEL): 19).

Burkhardt, Armin (1982): Gesprächswörter. Ihre lexikologische Bestimmung und lexikographische Beschreibung. In: Wolfgang Mentrup (Hg.): Konzepte zur Lexikographie. Studien zur Bedeutungserklärung in einsprachigen Wörterbüchern. Tübingen: Niemeyer (Reihe Germanistische Linguistik), S. 138–171.

Burnard, Lou (Ed.) (2007): Reference Guide for the British National Corpus. Available online at http://www.natcorp.ox.ac.uk/docs/URG/, checked on 5/1/2014.

Diewald, Gabriele (2006): Discourse particles and modal particles as grammatical elements. In: Kerstin Fischer (Hg.): Approaches to discourse particles. 1. Aufl. Amsterdam, Heidelberg: Elsevier (Studies in pragmatics), S. 403–425.

Duden. Die Grammatik: unentbehrlich für richtiges Deutsch (2006). Mannheim: Dudenverlag (Duden in zwölf Bänden, 4).

Engel, Ulrich (2004): Deutsche Grammatik. Neubearbeitung. Heidelberg: Groos.

Godfrey, J. J.; Holliman, E. C.; McDaniel, J. (1992): Switchboard: Telephone speech corpus for research and development. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 517–520.

Gohl, Christine; Günthner, Susanne (1999): Grammatikalisierung von weil als Diskursmarker in der gesprochenen Sprache. In: Zeitschrift für Sprachwissenschaft 18 (1). DOI: 10.1515/zfsw.1999.18.1.39.

Günthner, Susanne (2005): Grammatikalisierungs-/Pragmatikalisierungserscheinungen im alltäglichen Sprachgebrauch. Vom Diskurs zum Standard? In: Eichinger, Ludwig M. und Kallmeyer, Werner (Hg.): Standardvariation. Wie viel Variation verträgt die deutsche Sprache? Berlin, New York: De Gruyter, S. 41–62.

Helbig, Gerhard (2011): Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht. Unter Mitarbeit von Joachim Buscha. [Neubearb.]. Berlin, München, Wien, Zürich: Langenscheidt.

Hentschel, Elke; Weydt, Harald (2002): Die Wortart "Partikel". In: David A. Cruse (Hg.): Lexikologie. Lexicology, Bd. 2. 2 Bände. Berlin [u.a.]: De Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 21), S. 646–653.

Hirschmann, H., Lestmann, N., Rehbein, I., and Westpfahl, S. (2013). Erweiterung der Wortartenkategorien des STTS im Bereich 'ADV' und 'PTK'. Presentation at STTS Workshop, Hildesheim, Germany.

Hoffmann, Ludger (2013): Deutsche Grammatik: Grundlagen für Lehrerausbildung, Schule, Deutsch als Zweitsprache und Deutsch als Fremdsprache. Berlin: E. Schmidt. Online verfügbar unter http://deposit.d-nb.de/cgi-bin/dokserv?id=4057806&prov=M&dok%5Fvar=1&dok%5Fext=htm.

IDS, Datenbank für Gesprochenes Deutsch (DGD2). Online verfügbar unter http://dgd.ids-mannheim.de:8080/dgd/pragdb.dgd_extern.welcome?v_session_id=, zuletzt geprüft am 04.07.2014.

Imo, Wolfgang (2012): Wortart Diskursmarker? In: Björn Rothstein (Hg.): Nicht-flektierende Wortarten. Berlin: De Gruyter (Linguistik, Impulse & Tendenzen, 47), S. 48–88.

Institut für deutsche Sprache (2013): grammis 2.0. das grammatische informationssystem des instituts für deutsche sprache (ids). Unter Mitarbeit von Marek Konopka, Jacqueline Kubczak, Roman Schneider, Bruno Strecker, Eva Breindl-Hiller, Elke Donalies et al. Online verfügbar unter http://hypermedia.ids-mannheim.de/index.html, zuletzt geprüft am 17.07.2013.

Leuschner, Torsten (Hg.) (2005): Grammatikalisierung im Deutschen. Berlin: De Gruyter (Linguistik - Impulse & Tendenzen).

Oosterdijk, Nelleke (2000): The spoken Dutch corpus. Overview and first evaluation. Proceedings of the Second International Conference on Language Resources and Evaluation (LREC).

Rehbein, Ines; Schalowski, Sören (2013): STTS goes Kiez – Experiments on Annotating and Tagging Urban Youth Language. In Journal for Language Technology and Computational Linguistics (28(1)), pp. 199–227, checked on 4/30/2014.

Sampson, Geoffrey (2000): CHRISTINE Corpus: Documentation. University of Sussex. Available online at http://www.grsampson.net/ChrisDoc.html, updated on 8/18/2000, checked on 5/1/2014.

Schiller, Anne; Teufel, Simone; Stöckert, Christine; Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS. (Kleines und großes Tagset). Universität Stuttgart, Institut für maschinelle Sprachverarbeitung; Universität Tübingen, Seminar für Sprachwissenschaft. Online verfügbar unter http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf, zuletzt geprüft am 26.02.2014.

Telljohann, H.; Hinrichs, E.; Kübler, S.; Zinsmeister, Heike (2012): Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). University of T¨ubingenre.

Traugott, Elizabeth Closs (Hg.) (1997): The discourse connective "after all": A historical pragmatic account. 10th International Congress of Linguists. Paris, July.

VOICE (2013): Part-of-Speech Tagging and Lemmatization Manual. With assistance of Barbara Seidlhofer, Stefan Majewski, Ruth Osimk-Teasdale, Marie-Luise Pitzl, Michael Radeka, Nora Dorn. The Vienna-Oxford International Corpus of English. Available online at http://www.univie.ac.at/voice/documents/VOICE_tagging_manual.pdf, checked on 5/1/2014.

Weinrich, Harald (2005): Textgrammatik der deutschen Sprache. 3. Aufl. Hildesheim: Olms.

Westpfahl, Swantje; Schmidt, Thomas (2013): POS für(s) FOLK - Part of Speech Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch. In: Journal for Language Technology and Computational Linguistics (28 (1)), S. 139–153, zuletzt geprüft am 16.04.2014.

Zifonun, Gisela (1997): Grammatik der deutschen Sprache: [Bd. 1-3]. Hg. v. Ludger Hoffmann und Bruno Strecker. Berlin [u.a.]: De Gruyter (Schriften des Instituts für Deutsche Sprache, 7).