Published in: Calzolari, Nicoletta; Choukri, Khalid; Declerck, Thierry; Goggi, Sara; Grobelnik, Marko; Maegaard, Bente et al. (Hg.) (2016): Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia. Paris: European Language Resources Association (ELRA), pp.1493-1499.

FOLK-Gold – A GOLD standard for Part-of-Speech Tagging of Spoken German

Swantje Westpfahl, Thomas Schmidt
Institut für Deutsche Sprache
R5, 6-13 68181 Mannheim, Germany
E-mail: westpfahl@ids-mannheim.de, thomas.schmidt@ids-mannheim.de

Abstract

In this paper, we present a GOLD standard of part-of-speech tagged transcripts of spoken German. The GOLD standard data consists of four annotation layers – transcription (modified orthography), normalization (standard orthography), lemmatization and POS tags – all of which have undergone careful manual quality control. It comes with guidelines for the manual POS annotation of transcripts of German spoken data and an extended version of the STTS (Stuttgart Tübingen Tagset) which accounts for phenomena typically found in spontaneous spoken German. The GOLD standard was developed on the basis of the Research and Teaching Corpus of Spoken German, FOLK, and is, to our knowledge, the first such dataset based on a wide variety of spontaneous and authentic interaction types. It can be used as a basis for further development of language technology and corpus linguistic applications for German spoken language.

Keywords: German spoken language, POS tagging, GOLD standard

1. Introduction

In this paper, we present a GOLD standard of part-of-speech-tagged transcripts of spoken German. It comes with guidelines for the manual annotation of transcripts of spoken data and an extended version of the STTS (Stuttgart Tübingen Tagset) which accounts for phenomena typically found in spontaneous spoken German. The GOLD standard was developed on the basis of the Research and Teaching Corpus of Spoken German, FOLK (Schmidt 2014), and is thus, to our knowledge, the first such dataset based on a wide variety of spontaneous and authentic interaction types. The GOLD standard will be used in the FOLK project itself to further improve the part-of-speech tagging of the corpus. It will also be made available to the research community as a resource for the development of language technology for spontaneous interaction data.

The paper is structured as follows: section 2 briefly recapitulates related work, section 3 outlines the method by which data for the GOLD standard was sampled and manually annotated. Section 4 provides some technical information about the format of the GOLD standard and on how it will be made available. Section 5 gives an outlook on further work which we plan to carry out on its basis.

2. Related Work

There are several **corpora for German spoken data**: The Berlin Map Task Corpus (Sauer 2015), the Hamburg Map Task Corpus (HZSK 2010), GeWiss (Gesprochene Wissenschaftssprache - corpus of spoken academic language) (Fandrych et al. 2014), KiDKo (KiezDeutschKorpus - corpus of German youth language) (Rehbein et al. 2014), Tüba-D/S (Tübinger Baumbank des Deutschen/Spontansprache - Treebank of German spoken language) (Universität Tübingen, Seminar für Sprachwissenschaft 2014) are all corpora with transcriptions of specific types of conversations. Only the last two of them are manually annotated with part-of-speech tags, and only in KiDKo was the data annotated part-of-speech tags specific to spoken language (Rehbein/Schalowski 2013). Tüba-D/S consists of transcribed and manually annotated data (360,000 tokens) from the Verbmobil project (Wahlster 2000). The data is tagged with the original STTS (Schiller et al. 1999), hence discourse markers, hesitation markers, other speech particles as well as other spoken language phenomena are not accounted for. In KiDKo, 66,043 tokens were manually annotated with tags of an extended STTS tagset which was developed in cooperation with the work presented here.[1]

For other languages, several corpora of spoken language exist. However, similarly to the situation for the German language, only a few of them are manually annotated with part-of-speech tags or have tagsets adapted to their needs. The most important ones are the CHRISTINE Corpus (Sampson 2000), the Spoken Dutch Corpus (Oostdijk 2000) and the VOICE Corpus (Seidlhofer 2009).

3. Development of the GOLD standard

3.1 Sampling

The GOLD standard consists of 145 transcript excerpts. They were extracted as a cross-section of the 2014 release of the FOLK Corpus (totalling about 100 hours and 1 million tokens, see Schmidt 2014). The sample was designed to be balanced between the following dimensions which we expect to influence POS distributions:

- regional variants (e.g. speakers from Bavaria, Northern Germany) vs. standard German
- formal communication (e.g. university exam talks) vs. informal conversation (e.g. coffee-table talk)
- highly interactive interaction (multi-party, many

[1] For the GeWiss corpus, POS tagging experiments related to the ones presented here have been carried out and will be presented in Fandrych et al. (in preparation).

Speaker	Layer	Data	Free translation / *comment*
DK	Transcript	den würd ich nehmen	That one, I would take
	Normalization	den würd**e** ich nehmen	
	POS tags	PDS VAFIN PPER VVINF	
JZ	Transcript	ja beeil dich mal	Yes, hurry up *PPER 'dich' is reflexive*
	Normalization	ja beeil**e** dich mal	
	POS tags	NGIRR VVIMP PPER PTKMA	
	Transcript	(0.83)	*pause of 0.83 seconds*
MT	Transcript	((lacht))	((laughs))
JZ	Transcript	ich will ja noch nach hause	I still need to get home *'ja' is a model particle*
	Normalization	ich will ja noch nach **H**ause	
	POS tags	PPER VAFIN PTKMA PTKMWL APPR NN	
SK	Transcript	((stöhnt)) jetz fängt er damit an	((sighs)) now he is starting that *PTKVZ is the separated prefix of 'anfangen'*
	Normalization	jetz**t** fängt er damit an	
	POS tags	ADV VVFIN PPER ADV PTKVZ	
DK	Transcript	gome[z]	Gomez *player's name*
	Normalization	**G**omez	
	POS tags	NE	
PL	Transcript	[joa]	yes *yes/well/uhm*
	Normalization	**ja**	
	POS tags	NGIRR	
CH	Transcript	mh komm hey	oh c'mon!
	Normalization	**hm** komm hey	
	POS tags	NGIRR VVIMP NGIRR	
PL	Transcript	buh pfui ne spaßbremse	oh my, a spoilsport *'buh' and 'pfui' are interjections*
	Normalization	buh pfui **ei**ne **S**paßbremse	
	POS tags	NGIRR NGIRR ART NN	
CH	Transcript	gibt_s [nich]	no way *literally: 'does not exist'*
	Normalization	gibt **e**s nich**t**	
	POS tags	VVFIN PPER PTKNEG	
SK	Transcript	[m]h was bist_n du für_ne ++++++	uh what kind of (…) are you? *plus signs represent unintelligible speech*
	Normalization	**hm** was bist **denn** du für **ei**ne	
	POS tags	NGIRR PWS VAFIN PTKMA PPER APPR ART UI	

Table 1: Example from FOLK_E_00021 (Adults playing a football manager game) – square brackets = overlapping parts of conversation / double round brackets = para-verbal or non-verbal behavior of speakers

overlaps) vs. "disciplined" interaction (dialogues with ordered turn-taking such as in telephone calls)

We pre-selected a set of communications that would give us a satisfactory balance across these dimensions (with, for instance, map tasks and biographical interviews in different areas of Germany to ensure full coverage of regional variation, public discussions as instances of standard, formal and disciplined interaction, a poker game as an instance of non-standard, informal and highly interactive interaction, etc.) and extracted random samples of 500 to 1000 words from the corresponding transcriptions. Thus, the GOLD standard is relatively balanced concerning the interaction types sampled in the corpus: 41.6% of the transcripts contain clearly non-standard, mostly regional language, 46.7% contain standard language, and in 11.8% of the data some speakers talk in standard High German whilst their conversation partners speak in a regional variant. Whether speakers speak in a regional variant or standard High German can also be judged by the normalization rates annotated in the transcripts. 54.2% of our data can be categorized as formal conversations whilst 45.8% of the data are rather informal conversations. We have a slight bias towards more disciplined conversation (59.6%) as opposed to more interactive conversations (40.4%). One can also observe that the more informal and interactive conversations have a much higher rate in overlaps of speaker turns than formal, disciplined interactions. The large majority of speakers are fully competent native speakers. Nevertheless, for means of comparison, 1.3 % of the data contain utterances of non-native speakers and 3% of the data contain children's speech. We made sure that we extracted complete speaker contributions in order to minimize artefacts in the form of syntactically incomplete structures. Individual sample sizes therefore vary slightly around 500 or 1000 words. The total size of the GOLD standard amounts to 99,762 tokens (9,136 types).

The example in Table 1 illustrates a transcript excerpt with a transcription of words in modified orthography

according to GAT (Selting et al. 2009), measured pauses and description of para-verbal behaviour (laughter). Transcripts are aligned to the corresponding audio recordings with timestamps every 1 to 7 seconds. For each transcribed word, a mapping to its standard orthographic form (manually corrected with the help of OrthoNormal, see Schmidt 2012) is provided. POS tagging (and lemmatization) as described below were carried out on the normalized word forms with all non-word items ignored.

3.2 Manual Annotation

The GOLD standard was manually annotated with part-of-speech tags, i.e. the transcripts were first automatically tagged with the STTS (Schiller et al. 1999) using TreeTagger (Schmid 1995) with the standard parameter file. It was then manually corrected in an iterative process in order to improve both the automated tagging and the consistency of the manual annotation. This enables it to account for typical phenomena of spoken German.

The first step was the analysis of three automatically tagged transcripts (11,029 tokens) to determine the most frequent errors in the automatic part of speech annotation with STTS and TreeTagger. We found that with the standard TreeTagger parameter file for German texts and the default STTS version, POS tagging accuracy was as low as 81.16% (Westpfahl/Schmidt 2013). The most common cause of errors were the various types of speech particles, which accounted for more than 50 per cent of all errors. At the same time, these errors revealed a need for an adaption of the tagset to spoken language phenomena. Hence, before we started annotation of the GOLD standard, we adapted the tagset to this end. Furthermore, we developed a rule-based post-processing to automatically tag forms to which only one tag can be assigned, especially for those phenomena typically found in spoken language. An example would be the hesitation marker "äh" which is always tagged NGHES, the tag for hesitation markers.

3.3 Adapting the Tagset

In a second step, we used a subset of the GOLD standard – a development set of 24,229 tokens – to improve the tagset and guidelines. The adaption of the tagset is based on an in-depth linguistic analysis of the differences in grammatical categories between written language and spoken language. These analyses are part of the PhD thesis "*Part-of-Speech Tagging for German Spoken Data* – An analysis of spoken language phenomena with respect to automatically annotating the FOLK Corpus (Research and Teaching Corpus of German Spoken Language) at the Institute for the German Language in Mannheim (IDS Mannheim)" (Westpfahl 2017 (in preparation)).

Table 2 gives an overview of the most important changes and extensions to the tagset as defined in the new version of the guidelines "STTS für gesprochene Sprache".

As can be seen, the newly introduced categories follow the hierarchical scheme of the STTS (Schiller et al. 1999, p. 4). In addition, speech particle categories are entirely based on distributional features. They are categorized according to whether they are dependent on syntactic constructions and also in which phrase constructions they might occur. Thus, we added a supercategory for those elements which can occur independently of any other syntactic construction (NG), one for those elements which

Area	Changes/ Expansion	Tags	Explanation
Standard tagset STTS for written and spoken language	no punctuation	None	see above
	completion of category	PIDS	substituting indefinite pronoun with determiner
	deletion of category	PAV	prepositional adverb belongs to the class of adverbs
	change to another super-category	PTKANT	answering particle (Ger. 'Antwortpartikel')
	sentence-internal particles which are also used in written language	PTKIFG	intensifying, focus and scalar (Ger. 'Gradpartikeln') particles
		PTKMA	modal particles (Ger. 'Modal- and Abtönungspartikeln')
		PTKMWL	particles in lexicalized multi-word lexemes
Spoken language phenomena	POS-tag for disruptions within a wordform	AB	disruption (Ger. 'Abbruch')
	POS tag for spelled 'letters'	SPELL	Spelling
	POS tag for unintelligible utterances	UI	unintelligable (Ger. 'uninterpretierbar')
Speech particles	sentence-independent elements	NGIRR	interjections, backchannel signals (Ger. 'Rezeptionssignale') and responsives
		NGHES	hesitation markers (Ger. 'Häsitationspartikeln')
		NGAKW	inflectives (Ger. 'Aktionswörter')
		NGONO	onomatopoeia
	sentence-external elements	SEDM	sentence-external discourse markers
		SEQU	sentence-external tag-questions

Table 2: Overview of the most important changes and extensions to the tagset

can only occur within a syntactic construction (PTK) and one for those which are dependent on a subsequent or preceding construction but cannot occur within one (SE). The general aim was to keep the classes mutually exclusive but to allow for an exhaustive classification as well. It was intended to be as specific as possible, e.g. in annotating the function of word forms rather than just their morphological form, and as coarse-grained as necessary in order to minimize pragmatic interpretation and resulting inconsistencies in manual annotation. Thus, some word classes had to be grouped together under one category, especially in cases where the annotation would have strongly relied on the interpretation of the annotator or where word forms actually are ambiguous, for example backchannel signals, responsives and interjections (NGIRR).

As an example, consider PLs "joa" in the example given in Table 1 (a group of men playing football manager). This is the German equivalent to a mixture of "yes", "well" and "uhm", an interjection as well as a backchannel signal or answer to JZ's request to hurry up because he needs to get home soon. Reactions of everybody else in the group are negative towards that comment and PL's "joa" is an answer to JZ's request to hurry up as well as an interjection of disapproval. Pragmatically ambiguous phenomena like these had to be grouped together under one category. We are aware that in POS-annotation of spoken language, the mixture of various linguistic levels in the POS-categories is a problem. We opted to minimize that problem by following a distributional approach as far as possible and annotating pragmatic information only in those categories which are sentence-independent and sentence-external, i.e. where distributional information is insufficient by definition.

An inter-rater-agreement test between two annotators (trained student research assistants who were both involved in the developing process of the rules and guidelines) showed that the guidelines and the extended tagset worked well for the manual correction of part of speech tags (Cohen's Kappa of .98, see Westpfahl 2014).

Furthermore, the tagset presented here is also compatible with the STTS_IBK, a tagset extension for the tagging of data of computer-mediated communication (Beißwenger et al., 2015). In an STTS working group, we made sure that comparable phenomena were tagged the same way. This way, comparative corpus studies between spoken corpora and corpora of computer mediated communication will be facilitated in the future.

3.4 Retraining with the Development Set

A third step was the retraining of the TreeTagger with the development set. We trained the TreeTagger on a training set of 19,696 manually annotated tokens and evaluated the result on a set of 5,017 tokens. Tackling the sparse data problem, we added a word list (Institut für deutsche Sprache 2014) to the training process. The word list contains the most frequent 100,000 types of the tagged main archive of DeReKo 2014 (German Reference Corpus) (Institut für deutsche Sprache) which consists of about seven billion tokens. This choice was based on our intuition that the most frequent word forms are more likely to be tagged correctly as they would have been frequent in the training process as well. However, unlike Rehbein et al. (2014), we chose to analyse this wordlist and to clean out noise before using it as input for the tagger.

A close analysis of the first 25,000 tokens showed, however, that some categories had been systemically annotated incorrectly. This analysis was the impulse for creating a rule-based correction of the whole dictionary. For example, most of the items which were lemmatized as 'unknown' were tagged incorrectly so we decided to exclude them from the dictionary. Also, in German the first person plural and the infinite forms of verbs are identical so we added one entry with VVFIN (finite) to all forms in the dictionary labelled infinite except for those which are built with verb particles. Moreover, we excluded all entries with punctuation because it was either correctly tagged punctuation, which does not occur in our data or segmentation errors prone to be tagged incorrectly. Concerning pronouns we found that nearly all pronouns marked as substitutive can have a counterpart in the attributive class and vice versa. Hence we completed the entries for both classes. Furthermore we developed complete word form lists for all classes which are considered closed word classes and added those to the dictionary, i.e. for prepositions, determiners, conjunctions, subjunctions, verbal particles, pronouns, but also adverbs and a range of speech particle classes e.g. tag question markers or hesitation markers, onomatopoeia or interjections, response particles or backchannel signals.

We retrained the TreeTagger varying the parameters in different manners and evaluated the output. Our best result was obtained by a retraining of the TreeTagger adding our 'cleaned' dictionary and the post processing rules. This resulted in an accuracy of 91.01%. Compared to the original parameter setting (which resulted, on this evaluation set, in an accuracy of 80.92% after applying post processing rules correcting the most frequent speech particles), every retraining with the new tagset categories yielded substantial improvements on the accuracy. Adding a dictionary in order to counteract the out-of-vocabulary problem led to moderate improvements (around 2.3%). The cleaning of the dictionary did not seem to have any significant impact on the accuracy (<0.02%).

As we found the tagging improved, we tagged the rest of the GOLD standard with this retrained tagger. This sped up the manual annotation of the remaining 75% of the GOLD standard considerably. We ran another inter-rater-agreement test in order to make sure that the quality of tag annotations was not diminished by the fact that a) the number of errors to be corrected had decreased so much, b) the guidelines had been updated and c) the new student research assistants had not taken part in the development process. The result of this inter-rater-agreement was a Cohen's Kappa of .97.

4. Results

Our GOLD standard now consists of nearly 100,000 manually annotated word forms with lemmas and part-of-speech tags. The tagset and the tagset guidelines were developed to account for the most frequent phenomena in spoken language. They can be used as reliable tools for exhaustive but mutually exclusive word class annotations. The consistency and accuracy of the annotations were regularly controlled with inter-rater agreement tests.

In order to develop an automated POS tagging we used about 90% of the GOLD standard data to retrain the TreeTagger yet again. We evaluated its performance with more training input on both a balanced sample of transcripts of fully competent native speakers and on the transcripts of children's speech and non-native speakers. The results can be seen in table 3.

Evaluation data	POS acc %	super POS acc. %	ND POS acc. %	ND super POS acc %
Evaluation set	94,09	96,45	94,20	96,49
Learner language set	92,35	95,94	95,94	96,13

Table 3: Performance of the tagger after re-training;
POS acc. = accuracy according to the full tagset
super POS acc. = accuracy on the POS super-categories
ND = no dummies, items annotated with placeholder dummies were excluded in the training process

On the one hand, we evaluated the general accuracy of the gold tags. On the other hand, we evaluated to what extent the main (or super) category, at least, was correct. Since the tagset is structured hierarchically one can still see, for example, whether a finite verb is tagged as a verb, even when it is tagged as an infinitive. If the verb "schreiben" (write) in a sentence like "er will schreiben" (he wants to write) is tagged as VVFIN (finite full verb) instead of VVINF (infinitive of a full verb) one would still be able to find it in a corpus query looking for verbs in general because the supercategory "verb" was assigned correctly.

Furthermore, we also wanted to see whether elements typical of spoken language have an influence on the tagging process, namely disruptions, stuttering or aberrant interjections. All these phenomena are marked with dummy placeholders in the normalization process and are therefore easy to filter out in the training process and easily tagged in a post-processing.

Moreover, we were interested in seeing whether segmentation according to inter-pausal units would make a difference to the success of the training process. In our transcripts, no punctuation is added; pauses longer than 0.2 seconds are measured precisely and not assigned to any speaker in order to avoid interpretation on how one speaker said anything or to whose speech the pause belongs. The default case in our data is therefore that the speaker's contributions are segmented into inter-pausal units with a minimum pause length of 0.2 seconds, and it is these inter-pausal units that are fed into the tagger as sentence equivalents.

Following on, we wanted to find out whether a variation in the input, i.e. segmentation only after pauses of 0.3s, 0.5s, 1s or 3 seconds would have an influence on the retraining and on tagging results respectively.

In Table 4 one can see that the best results were achieved by leaving out those items which are marked by dummy placeholders in the training process. Segmentation according to longer pauses had almost no impact on the performance on competent speaker's data. Ultimately, segmentation with 0.3 seconds pauses as a segment boundary delivered the best results for learners' speech. Overall, however, these different changes to the input only had a marginal impact on the performance of the tagging process.

Test	POS acc. %	ND POS acc %	Learner POS acc %	ND Learner POS acc %
Pause > 0.2s (original)	94,09	94,20	92,35	92,56
Pause > 0.3s	94,01	94,15	92,33	92,58
Pause > 0.5s	94,14	94,11	92,44	92,51
Pause > 1.0s	93,98	94,19	92,44	92,56
Pause > 3.0s	93,89	94,13	92,19	92,49

Table 4: POS accuracy after re-training with various segmentation input

5. Distribution

In its finalized version, the GOLD standard data consists of four annotation layers – transcription (modified orthography), normalization (standard orthography), lemmatization and POS tags, all of which have undergone careful manual quality control. Taking up the first speaker contribution of the example in Table 1 from above, the GOLD standard will thus provide the following information on the token level:

Transcript	den	würd	ich	nehmen
Normalization	den	würde	ich	nehmen
Lemmatization	d	werden	ich	nehmen
POS	PDS	VAFIN	PPER	VVINF

Table 5: Annotation levels

The transcripts as a whole are structured into individual speaker contributions (corresponding to inter-pausal units with a minimum pause length of 0.2 seconds, see above). Comprehensive metadata documentation for the speakers and the interactions is available so that subsets of the GOLD standard can be systematically extracted. These represent specific speaker or interaction types (such as: informal speech of male speakers from Northern Germany). Transcripts are time-aligned in segments of 1.0 to 7.0 seconds, and the corresponding audio excerpts are available through the DGD.

The transcripts are stored in the FOLKER/OrthoNormal XML format (Schmidt 2012, see figure 1). In that form, they can be viewed easily and manually edited with the OrthoNormal tool. The format is fully compatible with the upcoming TEI-based ISO standard "ISO 24624 – Transcription of spoken language" (Schmidt 2011) and can be easily transformed to other formats useful for automatic and/or manual processing of the data. We will make the data available in its original format, in a TEI/ISO version and in a tabular separated format which

is common for language technology and corpus linguistic applications (such as TreeTagger and CQP, respectively).

```
<contribution speaker-reference="DK" start="TLI_279"
end="TLI_280" id="c203">
    <time timepoint-reference="TLI_279" time="0.0"/>
    <w id="w587" pos="PDS" lemma="d" n="den">den</w>
    <w id="w588" n="würde" pos="VAFIN" lemma="werden">würd</w>
    <w id="w589" pos="PPER" lemma="ich" n="ich">ich</w>
    <w id="w590" pos="VVINF" lemma="nehmen"
n="nehmen">nehmen</w>
    <time timepoint-reference="TLI_280" time="0.855"/>
</contribution>
```

Figure 1: XML representation

Textual data (i.e. annotated transcripts and metadata) will be distributed under a licence suitable for public and academic use via a web page of the Archive of Spoken German:

http://agd.ids-mannheim.de/folk-gold.shtml

The corresponding audio data will be made accessible through the DGD, which requires a free registration for members of academic institutions.

6. Outlook

We expect this GOLD standard to be of interest to a larger community in speech, language and corpus technology and hope that by making it available we can foster research into automatic processing of spontaneous spoken language data. Besides (re)tagging FOLK and other corpora in the Database for Spoken German, our own plans for future work on or with the data include:

1) The development and application of robust guidelines for segmenting the data into meaningful units above the token level (utterances, intonation phrases and the like). This is the objective of a French-German collaboration (funded by ANR/DFG) involving partners from Lyon (CLAPI database, Bert et al. 2010) and Orléans (ESLO corpus, Baude&Duga 2011), starting in spring 2016. The POS tagging of the GOLD standard is relevant in two ways for this project: on one hand, we expect POS tags to provide useful information for the (manual or partly automatic) segmentation process. On the other hand, we expect the performance of POS taggers to improve once the data has been segmented into units comparable to the sentences of written language (see above).

2) Experiments with other taggers. We have established an error rate of roughly 5% as a baseline for the task of automatic POS tagging of spontaneous spoken German. This baseline was achieved using the TreeTagger. It remains to be explored if POS tagging precision can be significantly improved with other tagging mechanisms. Next on our list is an experiment with a CRF tagger.

3) Integration of a POS tagger for spontaneous spoken German into tools widely used by the research communities working with spoken data. A TreeTagger based tool with a parameter file trained on the GOLD standard will be integrated into the EXMARaLDA system (Schmidt/Wörner 2014) and be made available as a web service in CLARIN so that it can, among others, be used as a component of a WebLicht tool chain (Hinrichs et al. 2010).

7. Bibliographical References

Baude, Olivier; Duga, Céline (2011): (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste? In: Corpus 10, Varia, 99-118.

Bert, Michel; Bruxelles, Sylvie; Etienne, Carole; Mondada, Lorenza; Traverso, Véronique (2010): Grands corpus et linguistique outillée pour l'étude du français en interaction (plateforme CLAPI et corpus CIEL). In: Pratiques - Interactions et corpus oraux, 17-34.

Fandrych, Christian; Meißner, Cordula; Slavcheva, Adriana (Hg.) (2014): *Gesprochene Wissenschaftssprache: korpusmethodische Fragen und empirische Analysen*. Heidelberg: Synchron (Wissenschaftskommunikation, 9).

Fandrych, Christian; Meißner, Cordula; Sadowski, Sabrina; Wallner, Franziska (in preparation): Gesprochene Wissenschaftssprache - digital. Verfahren zur Annotation und Analyse mündlicher Korpora. Stauffenburg Verlag. To appear.

Hedeland, Hanna; Schmidt, Thomas (2012): Technological and methodological challenges in creating, annotating and sharing a learner corpus of spoken German. In: Thomas Schmidt and Kai Wörner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*. Amsterdam/Philadelphia: John Benjamins, pp. 25-46.

Hinrichs, Erhard; Hinrichs, Marie; Zastrow, Thomas (2010): WebLicht: Web-Based LRT Services for German. In: Proceedings of the ACL 2010 System Demonstrations. pages 25–29.

HZSK (Hg.) (2010): *HAMATAC - the Hamburg MapTask Corpus. Archived in Hamburger Zentrum für Sprachkorpora.* Version 0.3. Publication date 2010-09-16. Retrieved February 21, 2016, from http://hdl.handle.net/11022/0000-0000-6330-A.

Institut für deutsche Sprache (2014): *Korpusbasierte Wortformenliste DeReWo. DeReKo-2014-II-Main-Archive-STT.100000*. Mannheim. Retrieved February 25, 2016, from http://www.ids-mannheim.de/derewo.

Institut für deutsche Sprache (2014): *Das Deutsche Referenzkorpus DeReKo*. Retrieved April 14, 2015, from http://www.ids-mannheim.de/kl/projekte/korpora.

Janin, A.; Baron, D.; Edwards, J.; Ellis, D.; Gelbart, D.; Morgan, N. et al. (2003): The ICSI Meeting Corpus. In: *Acoustics, Speech, and Signal Processing*, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on, Bd. 1, vol.1, pp. 1364-1367.

Mieskes, Margot; Strube, Michael (2006): Part-of-speech tagging of transcribed speech. In*: Proceedings of the*

5th International Conference on Language Resources and Evaluation (LREC). Genoa, Italy, pp. 935–938. Retrieved January 01, 2015, from http://www.lrec-conf.org/proceedings/lrec2006/.

Oostdijk, Nelleke (2000): *The Spoken Dutch Corpus. Overview and first Evaluation*. Dept. of Language and Speech, University of Nijmegen.

Oostdijk, Nelleke (2013): *Part of speech tagging. The Spoken Dutch Corpus*. Retrieved April 14, 2015, from http://lands.let.ru.nl/cgn/doc_English/topics/version_1.0/annot/pos_tagging/info.htm.

Rehbein, Ines; Hirschmann, Hagen (2014): Towards a syntactically motivated analysis of modifiers in German. In: *Proceedings of the 12th Edition of the Konvens Conference*, Hildesheim, Germany, pp. 30-39.

Rehbein, Ines; Schalowski, Sören (2013): STTS goes Kiez - Experiments on Annotating and Tagging Urban Youth Language. In: *Journal for Language Technology and Computational Linguistics (JLCL) 28* (1), pp. 199-227. Retrieved November 05, 2014, from http://www.jlcl.org/2013_Heft1/8Rehbein.pdf.

Rehbein, Ines; Schalowski, Sören; Wiese, Heike (2014): The KiezDeutsch Korpus (KiDKo) Release 1.0. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14),* pp. 3927–3934. Retrieved February 25, 2016, from http://www.lrec-conf.org/proceedings/lrec2014/index.html.

Sampson, Geoffrey (2000): *CHRISTINE Corpus: Documentation*. Release 2, 18 August 2000. Retrieved October 12, 2015, from http://www.grsampson.net/ChrisDoc.html.

Santorini, Beatrice (1990): *Part-of-Speech Tagging Guidelines for the Penn Treebank Project*. University of Pennsylvania. Retrieved February 25, 2016, from http://repository.upenn.edu/cgi/viewcontent.cgi?article=1603&context=cis_reports.

Sauer, Simon (Hg.) (2015): *BeMaTaC. Ein tief annotiertes multimodales Map-Task-Korpus gesprochener Lerner- und Muttersprache*. Humboldt-Universität zu Berlin. Retrieved October 06, 2015, from http://u.hu-berlin.de/bematac.

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. Retrieved February 26, 2014, from http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf.

Schmid, Helmut (1995): Improvements in Part-of-Speech Tagging with an Application to German. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung. In: *Proceedings of the ACL SIGDAT-Workshop*. Retrieved February 26, 2014, from http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf.

Schmidt, Thomas (2011): A TEI-based approach to standardising spoken language transcription. In: *Journal of the Text Encoding Initiative (1)*. Retrieved February 25, 2016, from http://jtei.revues.org/142.

Schmidt, Thomas (2012): EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language. In: *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12),* pp. 236–240. Retrieved February 25, 2016, from http://www.lrec-conf.org/proceedings/lrec2012/pdf/529_Paper.pdf.

Schmidt, Thomas (2014): The Research and Teaching Corpus of Spoken German - FOLK. In: *Proceedings of the Ninth International conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, pp. 383-387. Retrieved February 25, 2016, from http://www.lrec-conf.org/proceedings/lrec2014/index.html.

Schmidt, Thomas; Wörner, Kai (2014): EXMARaLDA. In: Durand, Jacques / Gut, Ulrike / Kristofferson, Gjert (eds.): Oxford Handbook of Corpus Phonology, Oxford: University Press.

Seidlhofer, Barbara (2009): *VOICE. The Vienna-Oxford International Corpus of English (version 1.0 online)*. Unter Mitarbeit von Angelika Breiteneder, Theresa Klimpfinger, Stefan Majewski und Marie-Luise Pitzl. Retrieved October 12, 2015, from http://voice.univie.ac.at.

Universität Tübingen, Seminar für Sprachwissenschaft (2014): *Die Baumbank TüBa-D/S*. Retrieved October 10, 2015, from http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-ds.html.

VOICE Project (2013): *VOICE Part-of-Speech Tagging and Lemmatization Manual*. Retrieved May 06, 2014, from http://www.univie.ac.at/voice/documents/VOICE_tagging_manual.pdf.

Wahlster, Wolfgang (2000): *VERBMOBIL. Erkennung, Analyse, Transfer, Generierung und Synthese von Spontansprache*. Hg. v. Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI GmbH). Retrieved March 24, 2015, from http://verbmobil.dfki.de/Vm.Info.Phase2.html.

Westpfahl, Swantje (2014): STTS 2.0? Improving the Tagset for the Part-of-Speech-Tagging of German Spoken Data. In: Lori Levin und Manfred Stede, editors, *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, pp. 1-10. Retrieved October 22, 2015, from http://www.aclweb.org/anthology/W14-4901.

Westpfahl, Swantje (2017) (in preparation): *Entwicklung eines automatisierten Part-of-Speech-Taggings für deutsche spontansprachliche Daten am Beispiel des Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)*. Dissertation. Universität Mannheim.

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 (JLCL) 28 (1)*, pp. 139-153. Retrieved April 22, 2015, from http://www.jlcl.org/2013_Heft1/6Westpfahl.pdf.