Originally published in: Schmidt, Thomas/Wörner, Kai (Eds.): Multilingual Corpora and Multilingual Corpus Analysis. - Amsterdam: Benjamins, 2012. Pp. 25-46.
(Hamburg Studies on Multilingualism 14)
DOI: *https://doi.org/10.1075/hsm.14.04hed*

POSTPRINT

Hanna Hedeland
Thomas Schmidt[1]

Technological and methodological challenges in creating, annotating and sharing a learner corpus of spoken German

Abstract: This article discusses questions concerning the creation, annotation and sharing of spoken language corpora. We use the Hamburg Map Task Corpus (HAMATAC), a small corpus in which advanced learners of German were recorded solving a map task, as an example to illustrate our main points. We first give an overview of the corpus creation and annotation process including recording, metadata documentation, transcription and semi-automatic annotation of the data. We then discuss the manual annotation of disfluencies as an example case in which many of the typical and challenging problems for data reuse – in particular the reliability of interpretative annotations – are revealed.

1. Introduction

Creating and annotating spoken language corpora is a laborious and time-consuming task; recording authentic interactions, transcribing them, and adding analytic information to the transcriptions are all jobs which demand advanced methodological skills and specialised equipment. The resulting corpora are thus precious resources, and nobody will seriously deny that enabling the research community to share and reuse such corpora is, on the whole, a worthwhile objective.[2] Typically, however, sharing and reusing data is, from a researcher's point of view, a subordinate aspect in corpus creation and annotation – after all, he will be primarily interested in carrying out his own analyses on the data. Thus, although

1. We are grateful to Kim-Chi Hamze, Seçil Yusun, Yael Dilger and Fideniz Ercan who supported us as student assistants in different stages of the corpus creation.

2. "[M]any researchers would agree that it is a basic scientific responsibility to make data collected in a research project available to the research community, especially when the research was supported by public funds" (The LIPPS Group 2000: 134).

the idea of data sharing is gaining more and more ground in empirical linguistics, the question of how decisions in technological and methodological matters during corpus creation and annotation affect the resource's value for later use has, in our opinion, not been discussed in as much detail as it deserves.

Our main motivation for building the Hamburg Maptask Corpus (HAMATAC) described in this chapter was therefore to create a basis for discussing these questions. In this way, we gained the necessary insight into corpus creation processes and were able to both test our EXMARaLDA tools with various annotation scenarios and analyse different annotation tasks with a view to technological and methodological aspects of reusability. Our focus in this paper is thus not on HAMATAC as yet another corpus, but rather on technological and methodological questions relevant for the reusability of spoken language corpora in general, exemplified on HAMATAC.

These questions become important in the phase of corpus design and at the collection of relevant metadata on speakers and communication events. Similarity to other corpora can enhance reusability, but comparability to other corpora through metadata is a prerequisite for reusability. Later, the choice between existing software tools, data models and formats determines the flexibility and sustainability of the corpus data itself. Questions of reusability are also very important when developing transcription and annotation methods, schemas and guidelines. Since this phase has perhaps the most complex relation to reusability of corpus data it will be discussed more in depth.

The paper is structured as follows: Section 2 introduces the Hamburg Map Task Corpus as our object of study. We present some general properties of that corpus and motivate our choice of corpus design in Section 2.1, outline the main steps in the corpus construction workflow in Section 2.2, and finally introduce and describe the four typical annotation tasks we decided to include in HAMATAC in Section 2.3.

Section 3 then picks out one particular task in that workflow – the manual annotation of disfluencies. This type of annotation corresponds to a very common task in the projects we (as the EXMARaLDA team) have supported during the last decade: project-specific interpretative manual annotation. We thus discuss the entire process thoroughly. Departing from a review of relevant theoretical approaches (Section 3.1) that results in a first draft of an annotation scheme (Section 3.2), we explain the implementation of the abstract scheme for EXMARaLDA (Section 3.3) and describe the further development of annotation categories and corresponding annotation guidelines (Section 3.4) as well as their practical application (Section 3.5). Finally, we present some disagreement measurement figures and discuss how these and similar annotation results should be regarded when it comes to questions of reusability (Section 3.6).

In our general conclusion (Section 4), we summarize our experiences with HAMATAC and discuss transcription and annotation quality as a prerequisite for reusability of corpora in general.

2. The Hamburg Map Task Corpus

2.1 Corpus design

The Hamburg Map Task Corpus (henceforth: HAMATAC) was created between October 2009 and June 2011 in the project Z2 'Computer assisted methods for the creation and analysis of multilingual data' of the Research Centre on Multilingualism at the University of Hamburg. It consists of recordings of advanced adult learners of German solving a map task, a common experiment for eliciting quasi-spontaneous linguistic data.[3]

The choice of that particular type of data was motivated by several considerations:

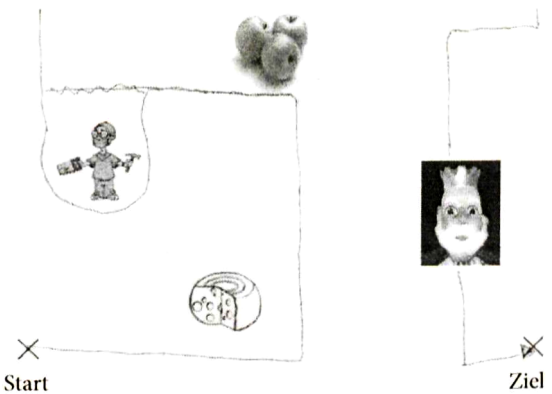

Figure 1. Excerpt of a map with a path drawn from start to finish (*'Ziel'*) by the instruction taker

3. "The Map Task is a cooperative task involving two participants. The two speakers sit opposite one another and each has a map which the other cannot see. One speaker – designated the Instruction Giver – has a route marked on her map; the other speaker – the Instruction Follower – has no route. The speakers are told that their goal is to reproduce the Instruction Giver's route on the Instruction Follower's map. The maps are not identical and the speakers are told this explicitly at the beginning of their first session. It is, however, up to them to discover how the two maps differ." [from the documentation of the HCRC map task corpus at http://www.hcrc.ed.ac.uk/maptask/maptask-description.html]

First and foremost, map tasks are interesting for our purposes (i.e. investigating efficiency and reusability of manual annotation) because they contain, in a relatively high density, many linguistic phenomena which lend themselves to systematic annotation. The disfluencies discussed in the following section are one such phenomenon – further phenomena comprise, for instance, learner "errors", backchannel mechanisms or prepositional phrases.

Second, map task recordings are relatively easy to obtain since they do not (in contrast to, for instance, many corpora in ethnomethodological research) require a specific environment or a specific field access.

Third, map task recordings are largely unproblematic in terms of data protection issues since they typically contain (in contrast to, for instance, recordings of private conversations) no or very little sensitive data – the persons recorded could therefore be easily persuaded to give their consent to a publication of the (anonymised) recordings.

Last but not least, HAMATAC's design provides interesting points of contact to other work in the fields of corpus technology and multilingualism research. More specifically:

- We use a map task designed in the project '*Variation des gesprochenen Deutsch*' at the Institute for German Language in Mannheim.[4] The '*Deutsch Heute*' corpus of this project (Brinckmann et al. 2008) documents regional variation of German. In a way, HAMATAC complements this corpus by recording learner (instead of native) variants of German, and this opens interesting opportunities for contrastive studies in the combined reuse of both corpora.
- The FALKO corpus from the Humboldt University Berlin (Lüdeling et al. 2008) is a corpus of written German learner language. The FALKO project has developed an extensive methodology for annotating errors in this corpus, and, again, FALKO and HAMATAC as written and spoken variants of learner corpora complement each other in a way which makes the potential benefits of data sharing obvious.
- The HCRC Map Task Corpus from the University of Edinburgh has been used as test and demonstration data for a number of corpus encoding techniques from the standoff annotation framework (Isard 2001). Our EXMARaLDA system[5] being a different, but related framework, the annotation of HAMATAC

4. http://www.ids-mannheim.de/prag/AusVar/Deutsch_heute/

5. EXMARaLDA is the system for creating, managing and analysing spoken language corpora which our project developed in the last ten years. A detailed description of the system's design and implementation can be found, for instance, in Schmidt/Wörner (2009).

can offer insights into strengths and weaknesses of different approaches to corpus encoding.
- As the corpus documents the linguistic performance of multilingual individuals, it also has important commonalities with most of the other corpora compiled at the Research Centre on Multilingualism.[6]

HAMATAC comprises recordings of 24 different speakers whose mother tongues cover a broad spectrum of languages, including Romance languages (French, Galician, Spanish), Slavic languages (Russian, Polish, Bulgarian) Persian languages (Dari, Farsi) and diverse languages from Non-Indo-European families (Turkish, Arabic, Chinese, Japanese, Thai, Vietnamese). Since speakers were selected and contacted from the immediate environment of student assistants in the project, most of them are between 17 and 40 years old and have a higher education. The length of recordings ranges from 3:31 to 21:52 minutes;[7] the total duration amounts to 3:17 hours. The orthographic transcriptions consist of altogether 21433 word tokens and 1277 word types.

2.2 Recording, metadata and transcription

We used two sets of maps and recorded speakers in pairs, switching the roles of instruction giver and instruction taker for the second set of maps. Recordings were made as WAV audio files in a university seminar room, using an M-Audio-Microtrack II recording device. During recording, speakers were instructed not to look at each other. The recording person (a student assistant) was told to intervene as little as possible during the task. However, some recordings contain short interventions of the recording person in which she clarifies some detail of the task. Speakers were asked to sign an agreement that the recorded data and a transcription thereof could be published on a password-protected website and used for research and teaching purposes. The agreement guaranteed the speakers that person names would be anonymised in the recordings and pseudonymised in the transcriptions and metadata. The speakers also had to fill in a short metadata form inquiring about some basic biographic facts. As illustrated in Figure 2, this data was entered as metadata in the EXMARaLDA Corpus Manager (see Wörner, this volume) and can be accessed alongside queries on the transcribed data using the EXAKT tool (Schmidt/Wörner 2009).

---

6. http://www.corpora.uni-hamburg.de/sfb538/en_overview.html

7. Longer recordings are an indicator of substantial communication problems between the respective participants.

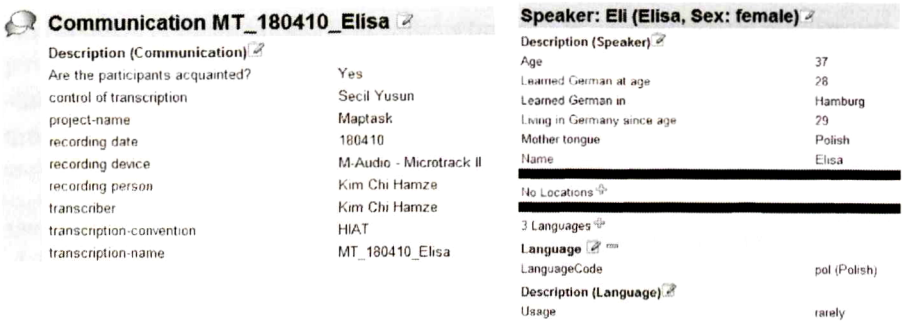

Figure 2. Excerpt of metadata for a communication (left) and a speaker (right) as displayed by the EXMARaLDA corpus manager

All recordings were transcribed by student assistants using the EXMARaLDA Partitur-Editor (Schmidt/Wörner 2009, see Figure 8 for a screenshot). They were instructed to use a simplified version of the HIAT transcription convention (Rehbein et al. 2004) concentrating on the following points:

- Words are transcribed orthographically without any punctuation in between. Capitalisation is used for nouns and proper names, but not at the beginning of a turn.
- Strong deviations from standard pronunciation and idiosyncratic forms are transcribed in 'literary transcription', e.g. *zweiderthalb* as an idiosyncratic from of *zweieinhalb*. Smaller deviations are 'orthographically corrected'. In particular, no attempt is made to represent foreign accents and similar phenomena inside the orthographic transcription.
- All pauses longer than 0.1 seconds are measured and attributed to the following speaker.
- Non-phonological productions (like laughing or coughing) are described in the speaker's tier between double round brackets, e.g. *((hustet))* for a cough.
- Cut-off words (typically in self-repairs) are marked with a slash.

After completion, transcriptions were double checked by at least one other person in the project.

2.3 Annotation

As mentioned in the introduction, we defined four different tasks which can be regarded as typical exemplars of distinct classes of corpus annotation in general – a disfluency annotation, a part-of-speech tagging, a lemmatisation, and an

dann	gehst	du	rechts	in	Richtung	äh	des	Rades	orthographic transcription
ADV	VVFIN	PPER	ADV	APPR	NN	ITJ	ART	NN	part-of-speech tag
dann	*gehen*	*du*	*rechts*	*in*	*Richtung*	*äh*	*d*	*Rad*	lemma
[dan]	[ge: st]	[du:]	[rɛ çts]	[ɪ n]	[rɪ çtʊ ŋ]	[ɛ :]	[dɛ s]	[ra:də s]	phonological form

Figure 3. Excerpt of an orthographic transcription with POS tags, lemmas and phonological annotation

annotation of phonological forms. These tasks differ in important characteristics along several dimensions:

The main difference between the disfluency annotation and the remaining three tasks is that they exhibit different degrees of interpretativeness. Assigning a lemma or a phonological form to some orthographically transcribed word is, in essence, a context-independent, form-based mapping based on intersubjectively available linguistic knowledge. By contrast, identifying and classifying disfluencies is often only possible on the basis of the annotator's subjective understanding (i.e. interpretation) of larger units of discourse. As will be discussed in more detail in Section 3, more interpretative annotation tasks are especially challenging with respect to the reusability of a corpus. Moreover, since interpretation requires human interpreters, the more interpretative annotation tasks hardly lend themselves to automation.

Automated methods can therefore only be applied in the part-of-speech tagging, the lemmatisation and the phonological annotation. A first important difference between these three tasks is that the first is done with probabilistic methods, while the latter two are lexicon-based. In the case of phonological annotation, lexicon-based lookup can be combined with rule-based methods for forms not present in the lexicon. A second difference is that part-of-speech tagging draws from a finite (and relatively small) set of annotation categories (i.e. the tags) whereas the number of annotation values in the other two tasks is potentially unlimited. This is highly relevant for the technical realisation of the annotation process because tools for manual correction of automatically produced annotations will have to take this difference into account in order to be efficient.

HAMATAC was POS-tagged and lemmatised using the TreeTagger (Schmid 1995) with the default German parameter file, trained on written newspaper texts. The data were first tokenized using EXMARaLDA's segmentation functionality which segments and distinguishes words, punctuation, pauses and non-phonological segments. Only words and punctuation were fed as input into the tagger in the sequence in which they occur in the transcription. The tagging results (POS tags and lemmas) were saved as EXMARaLDA standoff annotation files which can be further processed in the Sextant tool (Wörner 2010) and later integrated into the segmented transcription on separate annotation tiers.

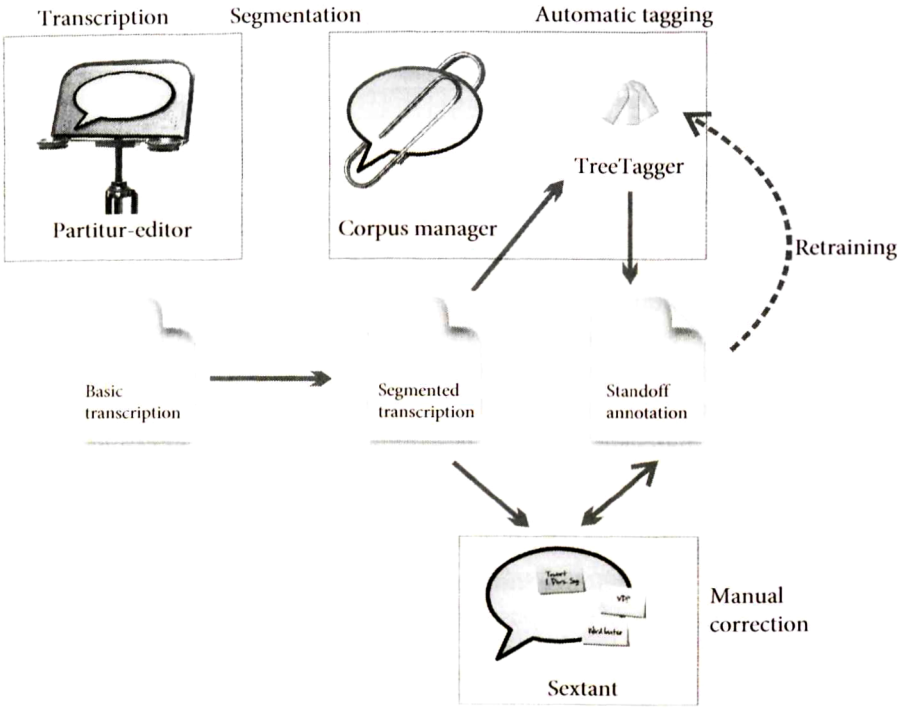

Figure 4. Workflow for POS annotation using TreeTagger

A student assistant was instructed to manually check and correct all POS tags using Sextant. An evaluation of 15 of the total 24 files shows that roughly 80% of POS tags were assigned correctly. The error rate is thus considerably higher than for the best results which can be obtained on written texts (about 97% correct tags). By far the most tagging errors, however, occurred with word forms which are specific to spoken language, such as hesitation markers (*"äh"*, *"ähm"*), interjections and incomplete forms (cut-off words). Since especially the former are highly frequent but very limited in form (three forms *äh*, *ähm* and *hm* account for about half of the tagging errors), we expect a retraining of the TreeTagger parameter file on the corrected data to lead to a much lower error rate.

The phonological annotation consists in a mapping of each orthographically transcribed word to its canonical phonological form. Although this does of course not take into account the actual phonetics of the word uttered – which would be useful information in its own right because many of the speakers have non-native accents – we still think that it can provide a lot of additional value to the corpus. Most importantly, it will allow queries to the corpus for different realisations of a target phoneme or a combination of target phonemes. Work on the phonological

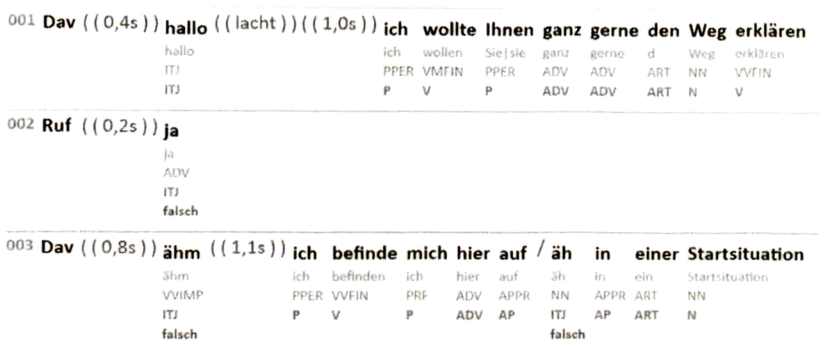

Figure 5. Visualisation of a POS tagged transcription with manual corrections

annotation of HAMATAC is ongoing. Like the part-of-speech tagging, phonological annotation presupposes the segmentation of the transcription into words. Our plan is to then look up each word form in HADIBOMP (http://www.sk.uni-bonn.de/forschung/phonetik/sprachsynthese/bomp), a large phonetic lexicon of German. If the word form is not present in the lexicon we will use an orthography to IPA conversion algorithm to calculate the most likely phonetic form.

3. Manual interpretative annotation

With the aim of testing the EXMARaLDA tools and gaining insight into this type of annotation process in general, disfluency seems a suitable topic for the manual annotation task. On the one hand, the recognition of disfluencies is an integral part of the widely used and adapted HIAT conventions and therefore already an issue during the basic transcription of the data. On the other hand, even a simple disfluency annotation scheme comprises categories that differ along the relevant dimensions of interpretation and representation. Both segments and points in the discourse require annotation; the annotated segments are sometimes words, sometimes longer arbitrary stretches of transcribed speech that the annotators have to detect and delimit; there are simple annotations and annotation sequences, sometimes including several levels and nesting; the categories are all interpretative but require different degrees of interpretation and the context to be considered for a single annotation also varies.

As stated in the introduction, we decided on a strict "bottom-up" approach. Instead of developing and promoting best practices "top-down", we tried to depart from – and, to a certain degree, imitate – the methods we have actually observed in many research projects. When imitating these methods, we attempted to

optimize the annotation process, but without imposing any of our own ideas of best practices in annotation.[8] One main property of such a "real life" annotation processes is that there is no systematic formalized method for quality control: Each transcription file is usually proofed – and corrected – by another project member, but disagreement in transcription or annotation is not systematically measured or evaluated. Some cases of disagreement considered to be particularly problematic or simply recurrent are discussed with all annotators to achieve agreement and thus reliable data. However, without systematic evaluation there is no well-founded knowledge about the resulting data. Through our investigation, we attempt to gain knowledge about what really happens in such processes, and how the choices made affect reusability.

3.1 Disfluency phenomena

Disfluencies in general and self-repairs in particular have been described in a similar manner by researchers with varying approaches and aims. We based our annotation scheme on McKelvie (1998), who models disfluencies to enhance parsing results, but also considered Levelt's (1983) classical study of self-repairs, and Hoffmann's (1991) discussion of anacoluthic constructions. Disfluencies are also described as part of the HIAT transcription conventions (Rehbein et al. 2004) used in or adapted for many of the projects we have supported.

McKelvie (1998) counts pauses, fillers, repetitions, speech repairs and fresh starts as cases of disfluency and distinguishes between hesitations "where (usually) non-lexical material is inserted into an otherwise normal utterance" and repairs "where some speech is retraced and later corrected" (1998: 10ff.). Hoffmann (1991) and Levelt (1983) both describe repetitions as a special case of repairs. We aim to also represent cascading or nested structures created by multiple repairs (Levelt 1983) and consider disfluency characteristics of words, i.e. lengthening, stuttering, word internal pauses etc. to render these words disfluent. In Figure 6, the essential similarities and differences of McKelvie's (1998) analysis (also representing Hoffmann (1991) and the HIAT conventions) and Levelt's (1983) are outlined.

Determining the segment boundaries of each different segment as represented in Figure 6 is obviously a difficult task when moving from simple examples to corpus annotation: Hoffmann (1991) refers to both reparandum (the part of the utterance to be repaired) and repair (reparans) as "segments", and also describes

8. Such suggestions for best practices in annotation do exist – a rigorous methodological approach for creating reliable annotation data can be found, for instance, in Bayerl et al. (2003).

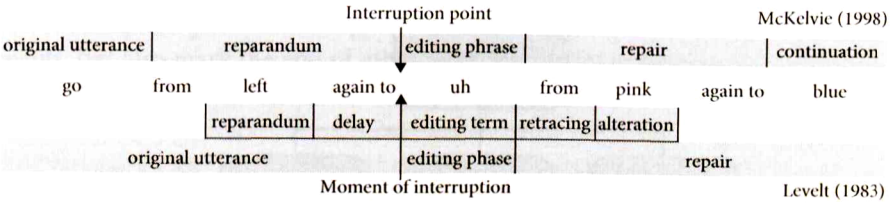

Figure 6. McKelvie's (1998) and Levelt's (1983) analyses of the structure of self-repairs

a common understanding of repairs that one could refer to as "the replacement idea": the replacement of the reparans, which is deleted from the utterance, with the reparandum. Still, he never marks the right boundary of the reparans(-segment) in his examples, thus completely avoiding this problematic aspect of disfluency annotation.[9] The HIAT conventions also only mark the interruption point, although they describe the same replacement idea (Rehbein et al. 2004). In conclusion, though both approaches agree on the existence of two segments that replace each other in a speaker's utterance, neither identifies their exact boundaries.

3.2 The HAMATAC disfluency annotation scheme

With use and potential reuse of the scheme in mind, one of the main desired features of our annotation scheme was simplicity. We wanted to make sure that student assistants could perform the annotation task as intended after some basic training. Another important aspect when outlining the rather basic categories was to include the replacement idea that also forms the basis for repairs in the widely used HIAT transcription conventions. Finally, we believe that segments should have to be identified if they do exist, even if this is a highly interpretative and very difficult task. While Levelt's (1983) more fine-grained analysis (see Figure 6) might be less interpretative and problematic on a lower level, the boundaries of the higher level's original utterance and repair rely on the highly problematic notion of "sentence boundaries", and so we decided against it. The HAMATAC Disfluency annotation scheme is presented in Figure 7.

Most of the categories pictured in Figure 7 describe segments:

- The TROUBLE category corresponds to the reparandum (McKelvie 1998; Hoffmann 1991), i.e. the part of the utterance that is to be deleted and replaced.

9. He argues that comprehension does not require the hearer to identify this boundary.

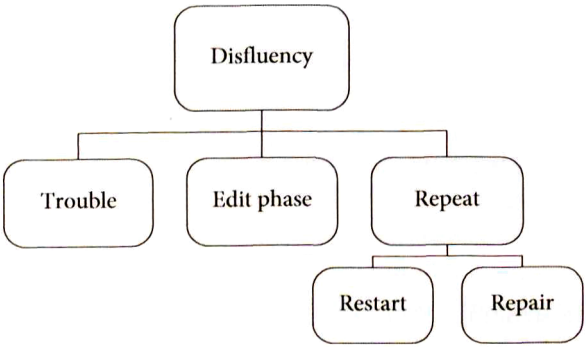

Figure 7. The HAMATAC Disfluency annotation scheme

- The REPAIR category comprises both repairs and reformulations (McKelvie 1998), i.e. any construction used to repair the TROUBLE part. To distinguish repairs from similar non-repairing structures, the guidelines require a recognizable interruption or an explicit meta-linguistic indication of the repair.
- The RESTART category is a repetition – it has retracing, but no alterations. Simple completion of words is also considered a RESTART.
- The REPEAT category is a common parent category for REPAIR and RESTART for cases where a decision is not feasible. Typically, the speaker is retracing, but, due to another interruption, it is not clear if alterations were intended. The choice between these categories (henceforth RE*-categories) is the only part of the annotation scheme for which the annotator has to choose between similar categories for a detected segment.
- The EDIT PHASE category is intentionally a very broad category, comprising all cases of delaying the formulation of an utterance, except for silent pauses, without retracing or looping, i.e. based on McKelvie's (1998) "hesitation". The following phenomena are classified as EDIT PHASE:
  - Filled pauses (usually *äh*, *ähm* or *hm*). Since we were annotating L2 German varieties, we chose not to restrict the inventory of sounds functioning as filled pauses a priori.
  - Words with phonetic characteristics that delay speech production, such as lengthening, within-word pauses or other signs of hesitation within words.
  - Meta-linguistic indications, i.e. words or phrases such as *nein* ("no"), *oder* ("or"), *ich meine* ("I mean") etc. used with repair signalizing function. No finite set of items has been defined.

The interruption points are obligatory points between TROUBLE- and RE*-segments, but also mark the end of other recognizable interruptions, for instance at aborted utterances, which are not further annotated.

3.3 Implementing the annotation scheme for EXMARaLDA

One interface between methodology and technology is the implementation of annotation schemes for the EXMARaLDA system. As discussed in (Schmidt 2005), time-based multi-tier formats allow for far more complexity than, for instance, inline tags and text or Word format.

As illustrated in Figure 8, the main annotation scheme was implemented using an annotation tier of the category "disfluency" for all segment categories described above. The extension of the individual segments in repairs is determined through replacement of the reparandum by the reparans. For each TROUBLE, the obligatory following RE*-category operates on the previous TROUBLE-annotated segment in the same tier. The extension of an EDIT PHASE was changed from being a phase, i.e. a sequence of segments that required an annotation, into one annotation per segment to allow for automatic comparison of annotations.

Whereas annotation tiers are used for disfluent segments, the point of interruption, which is not a segment, is represented as a plain slash in the main transcription tier. Words with phonetic characteristics qualifying as EDIT PHASE

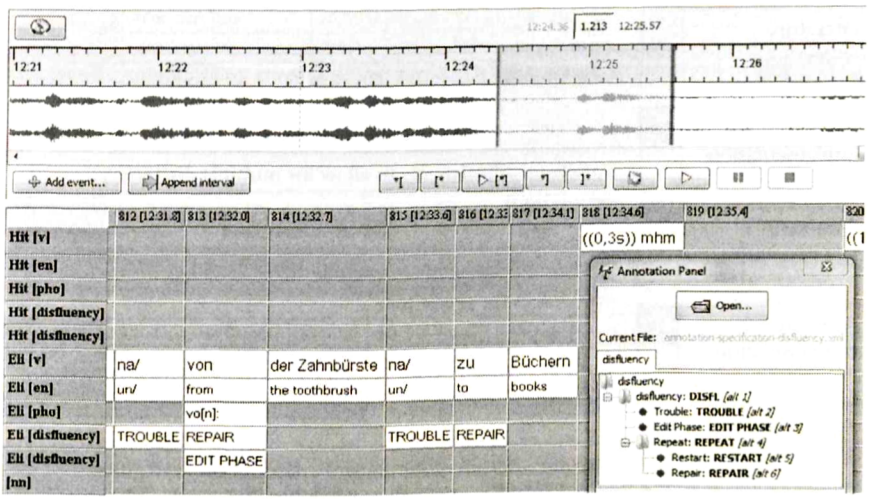

Figure 8. Using the HAMATAC Disfluency annotation scheme in the EXMARaLDA Partitur-Editor

were annotated with more detailed information in additional annotation tiers with the category "pho".

A reparandum can contain filled pauses, or a reparans might at the same time be a reparandum requiring another repair. Therefore, we needed to allow for multiple annotations of segments and adopted the straight-forward solution to allow multiple disfluency tiers. The tiers thus share the same tier category, but their unique tier-IDs can still convey more information on complexity and structure of a particular disfluency. Since annotation tiers are independent of each other, the EXMARaLDA format does not encode these structures as such. In fact, the interpretation that TROUBLE and REPAIR segments belong together is not explicitly encoded at all.

3.4 Developing annotation scheme and guidelines

Whereas the categories of the annotation scheme remained stable from the first draft, the guidelines with category definitions and practical instructions were developed in an iterative cycle (see Figure 9). In this phase, we worked closely together with two student assistants, both undergraduates studying for a Master in foreign language teaching/second language acquisition. First, one annotator spent one month testing the annotation scheme on the data, which resulted in many discussions of problematic cases. One of the main difficulties in creating

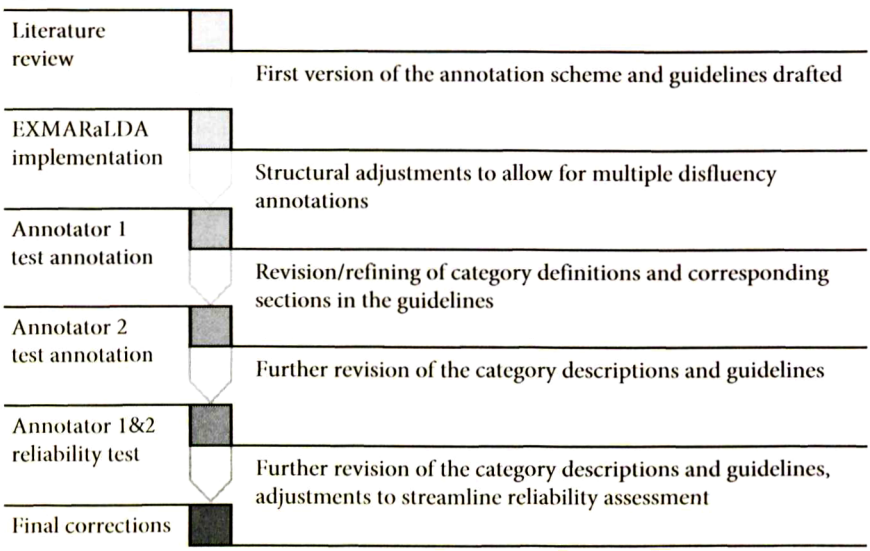

Figure 9. The annotation scheme development process

guidelines is that the exact boundaries of categories are often shaped by the cases encountered in the data and can therefore not be described a priori. The annotator contributed clarifications and examples for each category. As the second annotator joined the project, this process was iterated for two further weeks to make the guidelines more easily comprehensible. When both annotators felt confident about the annotation task, we conducted a reliability test with five transcriptions, about 4100 transcribed words, annotated independently by both annotators. The manual review of around 100 TROUBLE and RE* annotations, 250 EDIT PHASE annotations and 150 Interruption Points from each annotator – amounting to about 800 annotated segments or points detected by at least one of the annotators – lead to some final structural adjustments of the scheme to streamline future reliability assessment.

3.5 Using the HAMATAC disfluency annotation scheme

In this section, the application of the disfluency annotation scheme will be illustrated using authentic examples from HAMATAC.

Figure 10 shows a typical self-repair, comparable to the example in Figure 6. It illustrates how fluency can be achieved by replacing TROUBLE with REPAIR and removing all EDIT PHASEs (cf. McKelvie 1998).

Ali [v]	unter den	äh	((0,8s))	Nageln	((1s))	fährst/	öh	gehst	du nach oben
Ali [en]	below the	uh	((0,8s))	nails	((1s))	you drive/	uh	you go	upwards
Ali [disfluency]		EDIT PHASE				TROUBLE	EDIT PHASE	REPAIR	

Figure 10. Prototypic case of self-repair

Li [v]	eiter bisschen	da gibt es noch ne	Kro/	äh	Kro/	Kreise
Li [en]	a bit further	there, there's another	cirp/	äh	cirp/	circles
Li [disfluency]			TROUBLE	EDIT PHASE	RESTART	
Li [disfluency]					TROUBLE	REPAIR

Nad [v]	t	da wo du die Runde/	bevor du	die/	die	Runde	gemacht hast
Nad [en]		there where you turned/	before you	tur/	turned		around
Nad [disfluency]		TROUBLE	REPAIR				
Nad [disfluency]				TROUBLE	RESTART		

Hoa [v]	(und)	ma/	mache	ich dann/	schließe ich	den Kreis oben oder schließe ich den Kreis un	ten
Hoa [en]	(and)	d/	do	I then/	do I close	the circle at the top or do I close the circle at the bottom	
Hoa [disfluency]		TROUBLE			REPAIR		
Hoa [disfluency]		TROUBLE	RESTART				

Figure 11. Cascading, nesting and self-nesting repair structures

More complex structures such as those in Figure 11 motivated our use of multiple disfluency tiers. The examples also illustrate the difference between the RESTART and REPAIR categories.

3.6 Types and sources of disagreement between annotators

Since we chose (or imitated) an annotation process that is not systematically controlled, our aim in testing the reliability (cf. 3.4) was mainly to look into types and sources of disagreement in order to get a clearer picture of the interpretative manual annotation process. Out of 810 "cases" identified by at least one annotator, we had 489 cases where both annotators agreed on the category and the extension of the segment, 293 cases only identified by one annotator, 12 cases of different segment extension, 13 cases of differing categories and 3 cases of differences in both extension and category.[10]

The typology in Figure 12 illustrates the disagreement types discovered. We excluded "disagreement" due to fatigue or distraction and focused only on cases where the annotators agreed that they disagreed.

Some of the disagreement types will be illustrated below using examples from HAMATAC. There were 20 cases of type 1 disagreement, where one or both annotators made changes to the transcription. Figure 13 illustrates a case where the disagreement on the transcription (type 1c) related to the pronunciation of *wie* by a Vietnamese learner results in disagreement on the annotation category (type 2c).

1. Source on the transcription level:
   a. Characteristics of words
   b. Interruptions
   c. Wording
2. Source on the annotation level:
   a. Presence/absence of a simple annotation category
   b. Extension of a simple annotated segment
   c. Exact category of a simple annotated segment
   d. Structure of a complex repair sequence

Figure 12. Typology of disagreement

10. It might be tempting to compare such numbers with the corresponding figures of other annotation tasks. However, reliability metrics such as the ones developed by Carletta (1996) – although they avoid the drawbacks of raw percent agreement due to differences in number and distribution of categories – can only be applied when cases and the exact patterns of opposition between categories are defined before annotation. This is not the case with our approach.

Hit [v]	((1,4s))	m	((0,9s))	wie n/	((0,4s))	wie letzte	((0,5s))	unten	((1s))	also	Rad
Hit [pho]											[L]ad
Hit [en]	((1,4s))	m	((0,9s))	as n/	((0,4s))	as last	((0,5s))	below	((1s))	well	Wheel
Hit [disfluency]		EDIT PHASE		TROUBLE		REPEAT				EDIT PHASE	

Hit [v]	((1,4s))	m	((0,9s))	(wie)/	((0,4s))	(wie)	letzte	((0,4s))	unten	((1s))	also	Rad
Hit [pho]				wie[n]:		wie[n]						[L]ad
Hit [en]	((1,4s))	m	((0,9s))	(as)/	((0,4s))	(as)	last	((0,4s))	below	((1s))	well	wheel
Hit [disfluency]				TROUBLE		RESTART						

Figure 13. Disagreement type 1c (annotator A's "wie n/" vs. annotator B's "wie") causes disagreement type 2c (REPEAT vs. RESTART). Additionally, there are two cases of disagreement type 2a (only annotator A has EDIT PHASEs)

One might question our decision to allow the annotator to change the transcribed base of the annotation, arguing that true standoff annotation with an immutable base would have simply prevented these cases of disagreement. While this is true, the solution has a serious drawback: In forcing the annotators to work exclusively with the given transcription, one denies the most basic characteristics of transcription – its selectiveness and theory-dependence as well as the impossibility of ever declaring a transcription "complete" or "finished". Transcripts are always inherently interpretative and selective (Edwards 2001:321). With a focus on a certain linguistic phenomenon, the speech continuum might be interpreted differently since the conditions for the subjective perception have changed with the focus. The transcriber's and annotator's previous experience with learner language might also play an important role in their interpretation. This means there is no simple solution to this issue, although true standoff would have been a technologically pragmatic one.

The EDIT PHASE annotation of disfluent characteristics such as lengthening, while unproblematic in theory, turned out to be among the most problematic in the application.[11] It requires the annotator to divide a continuum into discrete categories, often dependent on global characteristics of the speech. This type of discretization is an important recurring issue in transcription – another example is the continuum between two related languages.[12]

Figure 14 illustrates a case of disagreement type 2d, where both annotators have detected repair sequences, but analyzed them differently. This kind of

11. In about 4100 words, annotator A detected 56 cases and annotator B 80 cases. Only 30 of the cases were detected by both annotators.

12. The general difficulty inherent in assigning words to a certain language is discussed by e.g. Gardner-Chloros et al. (1999:400).

Eli [v]	Nägel/	ähm	Ha/	m	((0,2s))	Nägel und Hammer
Eli [en]	nails	uh	ha/	m	((0,2s))	nails and hammer
Eli [disfluency]	TROUBLE	EDIT PHASE	REPAIR			
Eli [disfluency]			TROUBLE	EDIT PHASE		REPAIR

Eli [v]	Nägel	ähm	Ha/	mm	((0,2s))	Nägel und Hammer
Eli [en]	nails	uh	ha/	mm		nails and hammer
Eli [disfluency]		EDIT PHASE				
Eli [disfluency]	TROUBLE			EDIT PHASE		REPAIR

Figure 14. Degrees of disagreement in complex structures

Ali [v]	fährst/	öh	gehst	du nach oben/	((1,9s))	ziehst du nach oben	und	äh	ziehst rechts
Ali [en]	you drive/	uh	you go	upwards/	((1,9s))	you draw upwards	and	uh	draws to the right
Ali [disfluency]	TROUBLE	EDIT PHASE	REPAIR					EDIT PHASE	
Ali [disfluency]			TROUBLE			REPAIR			

Figure 15. In this continuation of Figure 10, annotator B recognized a repair where annotator A (Figure 10) saw a non-repairing reformulation ("gehst du" vs. "ziehst du")

disagreement on structured annotations is more complex than the trivial presence or absence of a category and difficult to describe and measure.

Two related issues lead to many non-trivial case discussions during the development process and after the reliability test: One was the application of the REPAIR category in ambiguous cases such as Figure 15. This is a highly interpretative task, since the annotator is forced to make a decision about the speaker's intention. And as McKelvie (1998: 11) points out, this decision also "depends on your theory of spontaneous speech" and the linguistic structures it allows.

The other issue was the replacement idea for TROUBLE and RE* segments. In cases where the relation between reparandum and reparans was not recognized as paradigmatic, category validity was questioned. That the relation is often more complex than mere replacement based on syntactic parallelism, has also been pointed out by McKelvie (1998) and Rehbein (1995). One might interpret this as a slight discrepancy between theory and supporting examples on the one hand, and the bulk of ill-fitting cases that do occur in spontaneous data on the other. The fundamental problem of this analysis seems to lie in the fact that it requires annotators to identify a larger construction, such as an utterance, that is fluent and well-formed after the replacement of reparandum with reparans. This is related to the forced exhaustive segmentation of spoken language common to most transcription systems, which has been questioned lately by Auer (2010).

4. Conclusion

As this paper has shown, decisions in every phase in the creation of a spoken language corpus can be said to have an impact on its later reusability. For instance, corpus design already co-determines possible reuses of the data through its relationships to the designs of other existing corpora. Also, practical issues such as an agreement to publish recordings have to be taken into account early in corpus construction in order to put data sharing on a sound legal base. Most importantly, however, reusability of a corpus is tightly linked to the question of how "reliable" transcriptions and annotations are.

We tried to make orthographic transcription – which, as the basis of all other annotation, is usually the most central task – more reliable by consciously reducing the complexity of the transcription convention. We also double checked all completed transcriptions. Still, as the disfluency annotation revealed, a need to modify smaller details of the transcription may still arise after several cycles of correction. As we have argued, this is an inherent property of spoken language transcription in general rather than an accidental characteristic of our particular corpus. This small source of unreliability should not be ignored when reusing a corpus, but we think that it is, on the whole, a minor problem.

The same holds for many types of annotation. As the POS tagging and lemmatization of HAMATAC have shown, existing technology allows us to automatically add several useful types of annotation to the transcription, and we can improve the quality of the results through manual correction with an acceptable investment of time and effort. The question of reliability only plays a minor role in this respect. It becomes crucial, however, when we turn to more interpretative annotation tasks.

The results from the disfluency annotation analysis point to the fact that in interpretative annotation tasks, annotators might never fully agree on all cases. Lampert and Ervin-Tripp even recognize "a tendency among coders from different backgrounds not to view and interpret language in exactly the same way, irrespective of the amount of training they receive" (Lampert/Ervin-Tripp 1993: 199). Therefore, guidelines can never replace annotators' implicit knowledge. It is not possible to explicitly and exhaustively state the required knowledge, and there is a limit to the amount of information annotators can internalize and make use of while annotating. A certain task might thus only be replicable with annotators who are similar with respect to their internalized knowledge – how to determine and describe this similarity is another difficult and seldom discussed question.

Due to the inherent ambiguity of human language and the interpretative nature of the annotation task, for some cases, different annotation solutions are equally possible and no single one is "correct". With current technology, it has

become feasible to encode different interpretations using a more complex corpus architecture, as the one implemented for the FALKO corpus (Lüdeling et al. 2008). On the other hand, allowing for conflicting interpretations does not imply that one annotator finds more than one, let alone an exhaustive listing, of interpretations.

Though these findings make the matter more complex, accounting for transcription and annotation quality remain a crucial prerequisite for data sharing. With inherently interpretative annotation categories, the degree of agreement on the interpretation determines how reliable the data is. Measuring reliability is however a highly complex matter that requires annotation scheme and task to be developed with this in mind. And even when reliability metrics can be applied, many questions remain unanswered. As thoroughly discussed by Artstein/Poesio (2008), reliability metrics aim to be comparable across annotation tasks by correcting for the expected agreement by chance resulting from number and distribution of categories in the annotation scheme. But to achieve comparability, other methodological and annotation scheme specific questions than those regarding the basic characteristics of annotation schemes need to be answered. Many of them are related to the choice of distance metrics for different types of weighted disagreement, for instance regarding related categories of annotation schemes[13], between unitizing and labeling disagreement for tasks including detecting and delimiting of segments for annotation[14] and to describe partial agreement for structured annotations (cf. Poesio/Artstein 2005). Another issue with comparability, the perceived difficulty of annotation tasks, can be illustrated by different cases of the EDIT PHASE category: Most filled pauses are manifest and thus easy to detect.[15] For lengthened words, on the other hand, discretization of a continuum is necessary (cf. 3.6). And to detect meta-linguistic signs of repairs, the contextual function of words must be interpreted. There are no agreed standards to describe these differences.[16] Thus, the results of reliability metrics remain difficult to interpret and compare across annotation tasks.

---

13. For the disfluency annotation scheme, disagreement between RESTART and REPAIR, which share the parent category REPEAT, should count less than disagreement between TROUBLE and REPAIR, and disagreement between REPEAT and one of its subcategories to count even less.

14. Is agreement on four out of five words labeled as a nominal phrase the same as agreement on four out of five words labeled as "loud"?

15. Both annotators detected 135 cases, only 6 cases were detected by only one annotator.

16. One alternative approach might be to measure the annotators' performance as Tomanek and Hahn (2010) did by recording time stamps for each annotated case.

To sum up, we think that most technological challenges in creating, annotating and sharing HAMATAC can be met in a satisfactory manner. Methodological challenges remain most of all in the area of highly interpretative annotation tasks. More work could be done here both on the theoretical side, especially with respect to the evaluation and measuring of reliability, and on the practical side, regarding workflows and best practices for the development and application of annotation guidelines.

References

Artstein, R. & Poesio, M. 2008. *Inter-coder agreement for computational linguistics.* Full version available online: http://cswww.essex.ac.uk/Research/nle/arrau/icagr.pdf, fetched 2011.04.21.

Auer, P. 2010. Zum Segmentierungsproblem in der Gesprochenen Sprache. In *LiSt* 49. <http://www.inlist.uni-bayreuth.de/> (November 2010).

Bayerl, P. S., Lüngen, H., Gut, U. & Paul, K. I. 2003. Methodology for reliable schema development and evaluation of manual annotations. In *Workshop notes for the workshop on knowledge markup and semantic annotation. Second international conference on knowledge capture (K-CAP 2003),* 17–23. Sanibel FL.

Brinckmann, C., Kleiner, S., Knöbl, R. & Berend, N. 2008. German Today: An areally extensive corpus of spoken Standard German. In *Proceedings 6th International Conference on Language Resources and Evaluation* (LREC 2008). Marrakesch, Marokko.

Carletta, J. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* 22(2): 249–254.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20: 37–46.

Edwards, J. A. 2001. The transcription of discourse. In *The Handbook of discourse analysis,* D. Schiffrin, D. Tannen & H. Hamilton (eds), 321–348. Malden MA: Blackwell.

Gardner-Chloros, P., Moyer, M., Sebba, M. & van Hout, R. 1999. Towards standardizing and sharing bilingual data. *International Journal of Bilingualism* 3(4): 395–424.

Hoffmann, L. 1991. Anakoluth und sprachliches Wissen. *Deutsche Sprache* 2: 97–120.

Isard, A. 2001. An XML Architecture for the HCRC Map Task Corpus. In *Proceedings of BI-DIALOG 2001,* P. Kühnlein, H. Rieser & H. Zeevat (eds). Bielefeld.

Krippendorff, K. 2004a. *Content Analysis: An Introduction to its Methodology,* 2nd edn. Thousand Oaks CA: Sage.

Krippendorff, K. 2004b. Measuring the reliability of qualitative text analysis data. *Quality and Quantity: International Journal of Methodology* 38: 787–800.

Lampert, M. D. & Ervin-Tripp, S. M. 1993. Structured coding for the study of language and social interaction. In *Talking data: Transcription and coding in discourse research,* J. A. Edwards & M. D. Lampert (eds), 169–206. Hillsdale NJ: Lawrence Erlbaum Associates.

Levelt, W. J. M. 1983. Monitoring and self-repair in speech. *Cognition* 14(1): 41–104.

The LIPPS Group. 2000. The LIDES coding manual. A document for preparing and analyzing language interaction data. Version 1.1 – July, 1999. *International Journal of Bilingualism* 4(2): 131–270.

Lüdeling, A., Doolittle, S., Hirschmann, H., Schmidt, K. & Walter, M. 2008. Das Lernerkorpus Falko. *Deutsch als Fremdsprache* 2: 67–73.

McKelvie, D. 1998. The syntax of disfluency in spontaneous spoken language [HCRC Research Paper HCRC/RP-95]. Edinburgh: Human Communication Research Centre.

Poesio, M. & Artstein, R. 2005. Annotating (anaphoric) ambiguity. In *Proceedings from the corpus linguistics conference 2005*. Birmingham: University of Birmingham.

Rehbein, J. 1995. *Segmentieren* [Memo 64]. Hamburg: Germanisches Seminar – Verbmobil (mimeo).

Rehbein, J., Schmidt, T., Meyer, B., Watzke, F. & Herkenrath, A. 2004. *Handbuch für das computergestützte Transkribieren nach HIAT* [Arbeiten zur Mehrsprachigkeit Folge B (Nr. 56)]. Hamburg: Sonderforschungsbereich Mehrsprachigkeit.

Reidsma, D. 2008. Annotations and subjective machines – of annotators, embodied agents, users, and other humans. PhD dissertation, University of Twente.

Schmid, H. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

Schmidt, T. 2005. Time-based data models and the Text Encoding Initiative's guidelines for transcription of speech. In *Arbeiten zur Mehrsprachigkeit* [Working Papers in Multilingualism, Serie B (62)]. Hamburg: Sonderforschungsbereich Mehrsprachigkeit.

Schmidt, T. & Wörner, K. 2009. EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics* 19: 565–582.

Siegel, S. & Castellan, J. 1988. *Nonparametric Statistics for the Social Sciences*, 2nd edn. New York NY: McGraw-Hill.

Tomanek, K. & Hahn,U. 2010. Annotation time stamps – Temporal metadata from the linguistic annotation process. In *LREC'10 - Proceedings of the Seventh International Conference Language Resources and Evaluation.*