A Syntax-Based Scheme for the Annotation and Segmentation of German Spoken Language Interactions

Swantje Westpfahl Institut für Deutsche Sprache Mannheim, Germany westpfahl@ids-mannheim.de Jan Gorisch Institut für Deutsche Sprache Mannheim, Germany gorisch@ids-mannheim.de

Abstract

Unlike corpora of written language where segmentation can mainly be derived from orthographic punctuation marks, the basis for segmenting spoken language corpora is not predetermined by the primary data, but rather has to be established by the corpus compilers. This impedes consistent querying and visualization of such data. Several ways of segmenting have been proposed, some of which are based on syntax. In this study, we developed and evaluated annotation and segmentation guidelines in reference to the topological field model for German. We can show that these guidelines are used consistently across annotators. We also investigated the influence of various interactional settings with a rather simple measure, the word-count per segment and unit-type. We observed that the word count and the distribution of each unit type differ in varying interactional settings and that our developed segmentation and annotation guidelines are used consistently across the social interactions that participants are involved in. This can be used for further analysis of social interaction and opens the possibility for automatic segmentation of transcripts.

1 Introduction

Since the beginning of research on spoken language, many different proposals for the segmentation of spoken language have been proposed. However, there is presently no segmentation system that could be used for large corpora of spoken language, i.e. a system that is linguistically substantiated as well as workable for large scale corpus segmentation. The lack of such a theory-based segmentation impedes the use of the corpora for research on language technology, comparative corpus linguistics as well as analyses in terms of spoken language interaction.

Focusing on the syntactic segmentation of such data has several advantages. Firstly, syntax theory is well understood in its application to written language corpora, so there are many tools for further processing of the data based on the scheme of syntactic units (e.g. POS, parsers, etc.). Secondly, a shallow syntactic segmentation and annotation can be used as a basis for linguistic analyses, also with various syntactic theories, pragmatic or prosodic approaches.

The need for a new segmentation of transcript data is based on the fact that the FOLK corpus (Schmidt, 2014b) in the DGD (data base for spoken German) (Schmidt, 2014a) is currently only segmented according to inter-pausal units, i.e. pauses longer than 0.2 seconds mark the boundary of each segment. This results in either very long sequences of speech when the speakers do not pause, or in segments of only partial structures when speakers make many pauses. This has two major disadvantages: One is the inconsistent visual representation that hinders researchers doing qualitative analyses based on transcripts, as it is common among conversation analysts, for example. The other is the limited value of such transcripts for contextually structured searches on the entire corpus, e.g. "Search for all instances of discourse markers at the beginning of a segment". Figure 1 shows an example of a very long speaker contribution that is merely segmented through pauses. Thus, the goal of a segmentation based on syntactic principles is that users of our corpus will be able to find whatever they are looking for in syntactic units rather than in inter-pausal units.

Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 109–120 Santa Fe, New Mexico, USA, August 25-26, 2018.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/

=	0106 FL	[mit einem]
≡	0107	(0.32)
=	0108 FL	r radius von zweihundertfünfzig meter für den es eine (.) ausnahmegenehmigung brauch was aber durchaus genehmigungsfähig is "h wo die güterzüge dann noch mit tempo achzig "h fahren könnten "h sie haben dann an beiden enden auch wieder die vorher schon besagten überwerfungsbauwerke da sie diese bahnstrecke "h kreuzungsfrei ausschleifen müssen "h insbesondere im bereich (.) der (.) remstalbahn vor dem (.) bahnhof (.) vor der es bahn haltestelle nürnberger straße "h ham sie dann den zustand dass dies sie die heute (sch) viergleisige trasse "h die in eimen engen einschnitt dort liegt (.) auf (.) sechs gleise aufweiten müssten
\equiv	0109	(0.21)
Ξ	0110 FL	°h was mit erheblichen (.) eingriffen in die (randbitter) bauung (.) äh ein einhergehen würde °h sie ham hier dann noch den bereich (.) mit einem gymnasium dem elly heuss knapp (.) gymnasium °h äh wo davon auszugehen wäre dass die überdeckung in diesem bereich des tunnels sehr gering wäre °h wo mer fragen muss ob das gebäude so (.) bestand haben könnte °h oder hinterher dann neu erstellt werden müsste

Figure 1: Example of the inconsistent visual representation of the data because of the current interpausal segmentation in the DGD, cf. Transkript FOLK_E_00064_SE_01_T_02_DF_01, 01:03:44.53 - 01:04:38.91.

In our study, we sampled a sub-corpus of various types of interactions. We developed a segmentation and annotation scheme based on the syntactic analysis of topological fields, i.e. the position of the finite verb and the complexity of the depending structures. We segmented and annotated this corpus according to this scheme. We validated the scheme and the reliability of the results with inter-annotator-agreement measures.

The research questions we aim at answering in this paper are: How reliable are segmentations and annotations on this kind of data, and where are the limits of syntax theory with transcripts of spoken language interaction? An additional question that arises naturally from the way we set up our test corpus is: How does syntax differ between interaction types?

2 Related work

As an approach for the descriptive analysis of the surface structure of the German syntax, the topological field model based on Drach (1937) has been widely discussed and further developed, e.g. by Wöllstein (2010; 2014) or Pittner and Berman (2013). In corpus linguistics, this model also has been used for the annotation of written data like the TüBa-D/Z (Telljohann et al., 2006), learner data like the Falco corpus (Reznicek et al., 2012), and to some extent also on transcriptions of spoken language (Andersen, 2008; Stegmann et al., 2000). In these works, the annotation of the topological fields is based on an existent segmentation of the data and is used as a basis for further linguistic analysis such as parsing as demonstrated e.g by Becker and Frank (2002). However, no solutions have been proposed for the annotation of typical spoken language phenomena as described below and thus their schemes for the annotation of topological fields cannot be used as a basis for the segmentation of transcripts of spoken language.

Since the development of oral corpora, the question and the need for a segmentation model of such data has been discussed. To date, segmentation into inter-pausal units is common for spoken language corpora as for example in the Switchboard corpus (Hamaker et al., 1998). However, for the analysis of syntactic structures a segmentation like this is not satisfying. Unlike sentences in written language data, transcripts of spoken interactions contain many features that do not appear in written language syntax. As Deppermann and Proske (2015) point out, many of the strategies for successful communication used by speakers is not sentence-like such as overlapping speech and collaborative turns, disruptions, ellipses and analepses, right and left dislocations, free topics, apokoinu constructions, vocatives and various types of speech particles such as backchannel signals, hesitation markers, and interjections.

Also Auer (2010) highlights the problems of transferring insights from research on written data to the analysis of spoken data and postulates four methodological principles considering the segmentation of corpus data. a) Exhaustiveness, i.e. the segmentation cannot leave out any material; b) Atomism, i.e. the segments must not include other segments of the same type; c) Discreteness, i.e. one element must

not be part of more than one segment, e.g. segment boundaries have to be clearly defined; d) Coherence of the linguistic level (Ebenenkonstanz), i.e. the segmentation has to be based on one approach of linguistic description. These principles must be considered in developing an approach for large scale corpus segmentation.

Many of the segmentation approaches proposed so far seem to work for selected examples or for specific types of data but violate at least one of Auer's methodological principles when it comes to the segmentation of entire corpora. However, some schemes have been developed specifically with the aim of corpus segmentation. In a pilot study, we assessed each of the following schemes (GAT2, HIAT, Macrosyntax) using our corpus data. Selting et al. (2009) presents a transcription system called GAT2 (Gesprächsanalytisches Transkriptionssystem) that is widely used in the field of Conversation Analysis (CA) for the representation of CA data extracts. According to GAT2, the segmentation of the data is based on the intonation phrase (IP), i.e. on prosodic cues. Problems with this approach arise from the fact that the definition for those IPs are circular in the GAT2 conventions and its criteria only vaguely defined as, e.g. in Grice and Bauman (2007) resulting in highly subjective interpretative annotations and segmentations. The circularity of these definitions can be explained as follows: the unit is defined by identifying "initial" or "final" movements in the intonation contour, i.e. the unit already has to be defined in order to identify the beginning and ending of it. Second, the definition gives several criteria for the identification of pitch movements pointing out that not all of them are necessary and some of them might be stronger or weaker in their presence. Thus, the definition remains rather vague and it is not clear which criteria must be fulfilled in order to constitute a unit (Selting et al., 2009, 370). We have encountered such problems in a test segmentation scenario in which three colleagues and experts on the field of conversation analysis have segmented the same transcript excerpts. The resulting segmentations turned out to be very different from one another.

With respect to pragmatic segmentation, two influential approaches have been proposed. The Handbook for computer-mediated transcription according to HIAT (Rehbein et al., 2004) relies on a method that translates into semi-interpretative working transcription (Halbinterpretative Arbeitstranskription) (Rehbein et al., 2004). The problem with this segmentation method is already implied in the title. It relies on subjective individual identification and interpretation of speech acts which proves to be problematic for a consistent analysis of the data especially with respect to inter-annotator agreements.

The second pragmatic approach is widely used in the segmentation of French oral corpora "Le protocol de codage macrosyntaxique". Strictly speaking, it is a mix of identifying illocutionary units with the help of the identification of structures according to dependency grammar (Benzitoun et al., 2012). Thus, this approach violates the principle of coherence of the linguistic level (cf. Auer (2010)). Moreover, first experiments on our German data with these guidelines have shown that in various cases the segmentation according to one linguistic level would contradict the other.

With respect to syntactic segmentation, there are two further approaches to be considered. The Analysis of Speech Unit (AS-unit) according to Foster et al. (2000) was developed for the segmentation and interpretation of English as a second language (ESL) and is based on the identification of syntactic structures taking into account typical spoken language phenomena. Yet, guidelines for the segmentation according to this approach are not published and there is no information on its performance with respect to inter-rater-agreements. The approach leaves the option for three levels of segmentation and only one level is supposed to segment the data exhaustively. Finally, it is advised that this kind of analysis should not be performed on data of very interactive interactions (Foster et al., 2000, 370f.).

Another approach for the segmentation according to a syntactic scheme was developed for texts of Early New High German where no punctuation is given and hence the segmentation of the data is necessarily done otherwise (Weiß and Schnelle, 2016). These guidelines describe several phenomena that also appear in transcripts of spoken language, such as parentheses and ellipses (Weiß and Schnelle, 2016). However, the most problematic cases for the segmentation of transcripts of spoken language such as disruptions, repetitions, and reported speech are not considered in that approach.

3 Material and method

3.1 Data and tools

The ultimate aim of our project is to find an approach to segmentation which can be applied on German as well as French and at the same time to analyze the language specific phenomena of each language. In order to be able to analyze various phenomena typical for spoken language interaction, it seems imperative to create a sample of various interaction types, which represent the variety of speech situations in our society (here, France and Germany). Hence, the pilot corpus comprises private interactions (e.g. family table talk) as well as public interactions (e.g. panel discussion or expert talk), informal faceto-face interactions (e.g. discussions with friends), and formal interactions, i.e. strongly governed by institutional rules (e.g. a school lesson or a panel discussion), speech influenced by written language (e.g. expert talk or reading a book to a child), and activity driven interactions (e.g. a cooking interaction or a service encounter), technically mediated communication (e.g. a telephone call between friends), and interactions with longer narrative passages (e.g. a biographical interview). Also, the number of speakers taking part in an interaction may vary from dyadic interactions to up to eleven participants. Moreover, the pilot corpus comprises data from speakers of the north, west, east, and south of Germany with some of them speaking rather standard high German and others speaking German influenced by their dialect. With this sampling we want to make sure that any segmentation scheme we develop is applicable for all of the data of our corpora in the DGD.

In sum, the German part of the pilot corpus that we are working on in this study comprises thirteen excerpts of transcripts from eleven interaction types. This amounts to about 110 minutes of audio material and about 22,150 tokens. The transcription is time-aligned with the audio data. For each transcription excerpt in German, there is an equivalent in French containing data of the same interaction type. The transcripts follow the cGAT transcription scheme (Schmidt et al., 2015) and are stored in the FOLKER format (Schmidt and Schütte, 2010). In order to change the segmentation and annotate the segments, we used the EXMARaLDA Partitur Editor (Schmidt, 2012).

3.2 Annotation and segmentation scheme

For the initial development of the segmentation guidelines we chose two transcripts: the telephone call and the panel discussion. They proved to exhibit many phenomena typical for spoken language interactions and also very different phenomena as the telephone call is a very interactive, private, and colloquial interaction whereas the panel discussion is structured by a moderator and the speakers can build very long and complex turns without anyone interrupting them. This way, we would already be faced with problematic structures typical for the different transcripts of spoken language in the development phase of the syntactic segmentation and annotation scheme.

Our first analyses of the data led to an inventory of segmentation problems, some of which are typical spoken language phenomena, some of them could appear in written language as well and especially so in computer mediated communication such as chat data or WhatsApp group communication. This inventory of segmentation problems can be grouped with respect to

- 1. specific phenomena on the word level such as discourse particles, hesitation markers, tag questions or vocatives,
- 2. characteristics of transcriptions themselves, such as transcribed breathing, nonverbal behaviour, vocal communication, pauses or alternative transcriptions, or
- 3. phenomena on the syntactic level such as disfluencies, disruptions, anacolutha, repairs, parenthesis, reported speech, expansions, lists with varying structures, ellipses, and collaborative turns.

Our syntactic segmentation and annotation scheme is composed of three annotation layers, which hierarchically depend on one another. The first annotation layer (cf. the second tier in Figure 2) is based on the identification of topological fields (Pittner and Berman, 2013; Wöllstein, 2014). The topological field model is based on the analysis of the surface structure of German sentences. One of the key features

of the German language is the verb bracket, i.e. the parts of the verb are anchored in two major positions. The left bracket constitutes either the first or the second position of the sentence, i.e. there can only be one constituent before the left bracket is realized. The right bracket is always at the end of the sentence. The topological field model is based on the identification of the verb brackets. The first constituent would then be the pre-field (**VF**, Vorfeld), followed by the left bracket (**LK**, Linke Klammer, e.g. a finite (auxiliary) verb), followed by the middle field (**MF**, Mittelfeld) in which the arguments of the verb and adjuncts are realized, followed by the (optional) right bracket (**RK**, Rechte Klammer, e.g. an infinitive, participle or verb particle). If material is added preceding this structure, which is also related syntactically to the sentence, it is situated in the pre-pre-field (**VVF**, Vor-Vorfeld), e.g. left dislocations, discourse markers, etc. If there is material added following the right bracket, this will be placed either in the post-field (**NF**, Nachfeld) in the case of right dislocations, i.e. material which would usually be placed in the middle field, or in the right outer field (**RAF**, Rechtes Außenfeld), e.g. question tags.

transcription	ja	also	dies	[.]	soll	ein	äh	beispielsatz	sein	weil	man	beispielsätze	braucht	ne	(0.24)	hey
field	KA	VVF	VF	LK		MF	1		RK	LK	MF		RK	NF		KA
clause	KVS	V2		-	VL											KVS
maximal unit	С	c											N			
translation	yes	well	this	(pause) is suppos	ed an	uhm	example	to be	becaus	e one	examples	needs	right		hey

Figure 2: Exemplary illustration of the segmentation and annotation scheme.

Our annotation scheme contains seven major categories for the annotation of topological fields as described, e.g. by Wöllstein (2010) or Pittner and Berman (2013). However, if the topological field cannot be specified because there is no verb in the utterance, we choose the annotation **KA** (keine Angabe - not specified). We also allow the possibility that the annotation of the topological field might be ambiguous (**AMB**) and that fields can be separated by parentheses in which case the field-annotations are numbered (...-1/...-2).

An annotation and segmentation scheme based on this model has several advantages. First, the identification of the syntactic structures relies on the surface structure of the text data, i.e. in the transcript, and explicitly not on characteristics that can only be found in the audio, e.g. prosodic features. This also helps to further process the data with automatic language processing tools that equally rely on the surface structure of the transcribed text. Second, the identification of the topological fields leads to the identification of larger syntactic structures, namely clauses. They also already provide the position of the finite verb and information on the type of clause that they occur in. In German, clauses with the finite verb in the second position are main clauses and mainly declarative sentences. If the finite verb is in the first position, e.g. if there is no pre-field, the clause is either an imperative, a question or sometimes a declarative with subject ellipsis, typical for colloquial speech (Auer, 1993). In German subordinate clauses, the finite verb is always in the last position.

In our annotation scheme, there are four major categories for clauses that contain a finite verb, the second of which has been newly introduced in our annotation scheme (cf. the second annotation layer (3rd tier) in Figure 2): V1 for verb first clauses, V1/2 for colloquial verb first clauses (e.g. subject ellipses), V2 for verb-second clauses, and VL for verb-last clauses.

Moreover, we added a category for the typical spoken language phenomenon of apokoinu constructions (**APO**). Yet again, if there is no finite verb in the utterance, an annotation according to the topological field model is not possible. Thus, we added two categories that are based on pragmatic linguistic analyses; The first one is for utterances that are complete in their pragmatic function, yet do not yield a finite verb (**KVS**, Kein Verb aber satzwertig). The second category is for structures that do not contain a finite verb and that do not fulfil the function of a sentence either (**KVN**, Kein Verb und nicht satzwertig), i.e. anacolutha. Also on this layer, the clause might be interrupted by a parenthesis, in which case we number the clause-annotations (...-1/...-2). Thus, our annotation scheme allows an exhaustive annotation of the data.

The annotation of the clause types has the advantage that the relation between various clauses can be

made visible, yet no information about their content gets lost. The dependencies between the various clauses are accounted for in the third annotation layer (cf. fourth tier in Figure 2), which at the same time represents the final segmentation of the data, i.e. the annotations of the maximal syntactic units. The categorisation of the maximal syntactic units represents the information gathered on the other levels of annotation and results in four categories (**S**, **C**, **N**, and **A**), which are specified as follows:

- **S:** The simple sentential unit consists of one and only one V1, V1/2, V2 or in very exceptional cases VL without any dependencies, cf. Figure 3.
- **C:** The complex sentential unit consists of several clauses that are dependent on one another: Main clauses with subordinate clauses or relative clauses, conditional sentences, reported speech, and matrix-clause with sentient-verbs, complex pre-pre-fields with main clause, discontinuous sentences, and coordinated sentences if and only if the second sentence shows subject or verb ellipsis, cf. Figure 2.
- N: Non-sentential units are all units that are not structured by a finite verb. On the clause level they are either annotated as KVS or KVN such as: (Sequences of) interjections, responses, or reception signals, words, and phrases without a finite verb, e.g. nominal phrases or prepositional phrases, or vocal communication, non-verbal behaviour, unintelligible utterances or vocatives, respectively, cf. Figure 4.
- A: An utterance which is disrupted, i.e. it opens a projection that is not fulfilled in what follows. This is segmented and tagged as abandoned unit (A), cf. Figure 5.

3.3 Segmentation problems and solutions

The categorisation already reflects some of the segmentation problems mentioned above. In addition, our guidelines provide additional rules for the handling of typical spoken language phenomena as well as for phenomena related to the transcription of spoken language. Thus, problems related to the groups (1.) and (2.) specified above, can be handled by simple rules that do not ground on any theory but that are rather formulated bearing in mind the ease of annotation and segmentation for the annotators.

Discourse particles such as interjections, response particles or reception signals always have a pragmatic meaning in context. Hence, they are segmented independently on the field and clause level, however, for matters of representation, they are subsumed on the maximal syntactic unit level as the initial "ja" in Figure 2 shows. Only when they occur surrounded by speaker pauses, they get their own, unspecified segment and are considered as non-sentential units, i.e. receive the tag "N" as the "hey" in Figure 2 shows. The example in Figure 3 illustrates the different handling of the discourse particle *ja* at the beginning and the transcribed breathing hh° at the end of the utterance.

EG [tok] [v]	ja	un	dann	ham	wir die feu	erwehr	angerufen	hh°	
EG [tok] [Feld]	KA	VVF	VF	LK	MF		RK		
EG [tok] [POV]	KVS	S V2							
EG [tok] [Max]	S								

Figure 3: Example for the annotation of discourse particles at the beginning of a turn.

In contrast, hesitation markers, if they co-occur with other parts of an utterance, are always segmented **within the following** segment as the "äh" in Figure 2. If the element is uttered at the end of a turn, it is segmented together with the preceding segment. Again, only when it occurs surrounded by speaker pauses, it gets its own, unspecified segment and is considered as non-sentential unit, i.e. receives the tag "N" as in Figure 4. The same rule holds true for transcribed breathing or micro pauses and any material that does not necessarily bear any semantic or pragmatic content.

Disruptions and self-corrections are a much bigger problem as they quite often result in ambiguous syntactic structures. In many cases, if at all, disambiguation is only possible with the help of context

SF [tok] [v]	äh
SF [tok] [Feld]	KA
SF [tok] [POV]	KVN
SF [tok] [Max]	N

Figure 4: Example for the annotation of hesitation markers surrounded by silence.

EG [tok] [v]	ja	wie	er	das	hat	fallen	lassen	wei	wieso	hast	du	das	paket	eigentlich	fallen	lassen
EG [tok] [Feld]	KA	VF						LK	VF	LK	MF				RK	d
EG [tok] [POV]	KVS	VL	VL V2 V2													
EG [tok] [Max]	A								S							
EG [tok] [v]	ja	wie	er	das	hat	fallen	lassen	wei	wieso	hast	du	das	paket	eigentlich	fallen	lassen
EG [tok] [Feld]	KA	VF						KA	VF	LK	MF		RK			
EG [tok] [POV]	KVS	S VL							T V2							
EG [tok] [Max]	S	5							S							

Figure 5: Example of an ambiguous case. The annotation guidelines indicate to take the lower option.

knowledge and/or knowledge about the prosody of the utterance as illustrated for the two annotation and segmentation versions of the example in Figure 5. By listening to the audio, one can guess that the disruption *wei* might have been the word *wei* β (know), which would result in the interpretation of a verb second clause preceded by a relative clause (first option). The whole structure would then have to be interpreted as an abandoned unit. If one, on the other hand, assumes that the verb-last clause is an exclamation, it would already constitute a unit for itself (second option). Thus, the abandoned unit would consist of the *wei* only. The solution we propose for this type of problem is parsing the data from the left to the right. If an utterance can be considered complete on the syntactic level, it is annotated as such (second option). This solution bears in mind future work with other language processing tools, most of which follow the same scheme.

3.4 Evaluation methods

We hired two student assistants who segmented and annotated the entire pilot corpus each and independently according to the scheme described above. In order to evaluate the annotation scheme, we measured the inter-annotator-agreement on the two annotated data sets, i.e. we calculated the raw agreement and a modified kappa value with the help of the ELAN tool (Wittenburg et al., 2006). The raw agreement measures the percentage of exactly similar segmentation and annotations. The modified kappa value uses the same basis but takes chance agreement into account. For the evaluation, we excluded the transcript excerpts used for the development of the guidelines, i.e. the telephone call and the panel discussion (cf. Table 4). The results are shown in Table 4.

In order to evaluate the segmented and annotated data, we chose a method proposed by Grabar and Eshkol (2016): we counted the segments and the tokens in each segment. This way we can retrieve information about the annotated segments with respect to (i) their length depending on the annotation (here: label), i.e. the average number of tokens in the segment, (ii) whether the average length of the segments differs with respect to the interaction type, and (iii) whether the annotators chose the same annotations in the same transcript type, i.e. whether the length of the segmentation and their labelling vary significantly. We used a chi-square test to evaluate the amount of syntactic unit types per interactional setting. Further, we used a linear regression model to predict the number of tokens in a segment by its label, the type of interaction, and the annotator (as well as all possible interactions). This model allows to address three questions at once: (a) Do segments differ in the number of tokens with respect to their labels? (b) Do annotators differ in their overall segmentations and labelling? (c) Do the number of tokens in segments of a specific category differ with respect to the annotator? The results of the latter questions also add to the analysis of inter-annotator-agreement.

4 Results and discussion

First, we present the results of the inter-annotator-agreement. As can be seen in Table 4, the results on inter-annotator agreement of topological fields are on average 84% raw agreement and result in a Kappa value of 0.81, indicating almost perfect agreement. The raw agreement on the clause level (pov) and on the maximal unit level (max) is slightly lower, i.e. around 74% and a Kappa value of 0.67, and 78% and a Kappa value of 0.69 respectively, indicating substantial agreement.

Table 1: Inter-Annotator-Agreement calculations (raw agreement and kappa) according to annotation layers (field, pov, max) and various interaction types. Excluded are the telephone call and the panel discussion, as they were the annotation training data.

		Acadi eddi coline	lable lalf	Social meer	Conficuration Confiction	Etpent all	School less	Interview	A Control of the the	Service encal	ounter
metric	layer	\$°	tan 1	So	co	\$\$*	స్తు	The second	22	సి	mean
agr.	[field]	89.55	82.32	79.36	80.11	79.32	87.97	84.13	89.26	87.33	84.37
a X	[pov]	81.02	72.43	65.53	65.29	71.88	71.69	74.25	83.01	80.80	73.99
raw	[max]	83.80	82.42	80.24	61.69	72.79	79.05	76.49	79.48	85.74	85.74
Da l	[field]	0.87	0.78	0.75	0.77	0.75	0.85	0.81	0.87	0.85	0.81
kappa	[pov]	0.73	0.63	0.58	0.57	0.67	0.63	0.67	0.78	0.73	0.67
k	[max]	0.76	0.70	0.68	0.53	0.63	0.69	0.69	0.71	0.71	0.69

The individual agreement measures indicate that the conflictual interaction as well as the meeting in the social institution are the most problematic transcript excerpts.



Figure 6: Mosaic plot: standardized residuals from the chi-square test showing the dependencies of the number of segments of a certain category with the interaction type. The height of the tiles represents the relative number of segments of each category in each transcript. The width of the tiles represents the number of segments in the whole transcript compared to the others.

The chi-square test reveals statistically significant differences in the distribution of unit types in the various interaction types. Figure 6 shows a mosaic plot visualizing the residuals of the test. For example, non-sentential units (N) are relatively less frequent in the conflictual interaction, the expert talk, the

interview, and the panel discussion, but relatively more frequent in the social meeting and the family table talk. Most interactions mainly consist of non-sentential units (N) and simple sentential units (S) whereas the expert talk and the panel discussion have an increased number of complex sentential units (C) relative to their total number of segments.



Figure 7: Effect plot: dependencies of the annotated categories (labels), the segment length (number of tokens in a segment along the y-axes), the interaction type (along the x-axes), and the annotators (AB = student assistant 1; HS = student assistant 2).

The linear regression model reveals statistically significant differences relating segment labels (N, S, A, and C), number of tokens, interaction types, and annotators. Figure 7 shows the effect plot. We can see that segments annotated as non-sentential units (N) are shorter in the number of tokens (with an average length of 2.26) than abandoned units (A) (with an average length of 5.09), which are in turn shorter than simple sentential units (S) (with an average length of 7.35 tokens), which are again shorter than complex sentential units (C) (with an average length of 16.84 tokens). These differences in the length of the segments are statistically significant, which implies that the categories we chose actually represent different syntactic phenomena. Moreover, the results show that different interaction types indeed vary significantly in the length of their syntactic units. Complex (C) and simple (S) sentential units are significantly longer in the expert talk than in all other types of interactions except for the panel discussion. In the table talk, the two unit types (C and S) are significantly shorter than in all other interaction.

With respect to the annotations of the two annotators, one can see that only in the interview the annotations of the abandoned unit category differ statistically significantly from one another. Hence, one can see that even though we focused on this problem in writing the guidelines, the abandoned units seem to be a greater challenge to segment and annotate than other units. In a detailed qualitative analysis, we found that many cases in which the annotators disagreed with respect to abandoned units result in true structural and syntactic ambiguities. These results imply that we have to improve the guidelines in a way that default rules for ambiguous cases are offered. The segmentation and annotation of the gold standard will include these improvements, which will hopefully resolve the remaining disagreements.

5 Conclusion and perspectives

In this paper, we could show that syntactic units can be segmented and annotated fairly reliably, and that these annotations add valuable information to the data. Our evaluation showed that there are indeed statistically significant differences in the syntax used by speakers in different interactional settings. On the one hand, private interactions generally contain much more speech particles and other non-sentential segments than for example interactions in the public. Also, their sentential units such as simple and complex sentences are generally shorter than in other interactions. On the other hand, one can see that if the speakers have the floor, they tend to make much fewer but more complex and longer syntactic units and even abandoned units are significantly longer than in other interactions.

Generally, abandoned units, e.g. disruptions, self-corrections, and anacolutha are the greatest challenge for the segmentation and annotation of syntactic structures. In contrast to other annotation approaches of topological fields and syntactic structures (Stegmann et al., 2000), we explicitly do not exclude these typical spoken language phenomena from the annotation. Thus, our segmentation and annotation scheme allows for an exhaustive segmentation of the data and also respects the methodological principles of atomism and discreteness described by Auer (2010). It also reflects the coherence of the linguistic level insofar as it is made clear in the guidelines that syntax is the dominating level of analysis and pragmatic knowledge is only used if there is no finite verb for the analysis of syntactic dependencies.

We improved the guidelines and established a reference segmentation for the pilot corpus, which we present as a gold standard and which we will make publicly available at the end of the project. Apart from the syntactic annotations described above, the gold standard also contains additional pragmatic annotations of typical spoken language phenomena such as disruptions, non-verbal behaviour, collaborative turns, vocal communication, vocatives, reported speech, parentheses, and unintelligible utterances.

With the improved guidelines we will segment and annotate a larger pilot corpus of about five hours of transcriptions, consisting of other interaction types as well as similar ones in order to corroborate the findings stated above.

With this study, we could also corroborate findings of conversational analysis that social actions determine the use of the language depending on the conversational situation. Our data can help in designing language processing tools adjusted to varying interaction types.

Due to the large size of spoken language corpora, one of our aims is to automatize their segmentation or at least to be able to reduce manual effort in segmenting the data. A gold standard of segmented data on several syntactic levels can be a valuable help for the development of automated segmentation. We also make use of online annotation experiments (Schmidt and Westpfahl, submitted) for the annotation of the boundary status of pauses of various types. Further parameters such as prosody or non-verbal behaviour, POS-tags or statistics derived from interaction type, speech rate, overall distribution of pauses, etc. could be considered as well. Finally, we hope to take advantage of the outcome of our various annotations in a multi-factorial model.

The solutions presented in this paper are language specific for German and related V2-languages. At a later stage of our project Segmentation of Oral Corpora we aim at finding a shared solution with our French cooperation partners (Interactions, Corpora, Apprentissages, Représentations (ICAR), Université Lyon and Laboratoire Ligérien de Linguistique (LLL), Université Orléans) who work on the segmentation of French spoken corpora. A shared inventory of segmentation problems helps in finding similar means to solve problems and identifies structures for which rules have to be found language specifically. The analyses based on the comparison of a Germanic and a Romance language may reveal structures which are language independent and could be transferred to other languages as well. However, as each language has its own syntax, language specific rules will have to be found in order to do justice to each language.

Acknowledgements

We would like to thank the ANR and DFG for funding this work in the project "Segmentation of oral Corpora" (SegCor, DFG project number 281693063). We are grateful for the collaboration, assistance and support of Thomas Schmidt, Hanna Strub, Anton Borlinghaus, and Robert Owen Jones.

References

- Christiane Andersen. 2008. Topologische Felder in einem Korpus der gesprochenen Sprache: Probleme zwischen theoretischem Modell und Annotation. *Göteborger Arbeitspapiere zur Sprachwissenschaft*, 3.
- Peter Auer. 1993. Zur Verbspitzenstellung im Gesprochenen Deutsch. Deutsche Sprache, 23:193-222.
- Peter Auer. 2010. Zum Segmentierungsproblem in der Gesprochenen Sprache. InLiSt Interaction and Linguistic Structures, 49.
- Markus Becker and Anette Frank. 2002. A stochastic topological parser for German. In *COLING '02*, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Christophe Benzitoun, Frédéric Sabio, Paola Pietrandrea, and Sylvain Kahane. 2012. Protocole de codage macrosyntaxique.
- Arnulf Deppermann and Nadine Proske. 2015. Grundeinheiten der Sprache und des Sprechens. In Christa Dürscheid and Jan Georg Schneider, editors, *Handbuch Satz, Äußerung, Schema*, 4, pages 17–47. De Gruyter, Berlin.
- Erich Drach. 1937. Grundgedanken der deutschen Satzlehre. Diesterweg, Frankfurt am Main.
- Pauline Foster, Alan Tonkyn, and Gillian Wigglesworth. 2000. Measuring Spoken Language: A Unit for All Reasons. *Applied Linguistics*, 21(3):354–375.
- Natalia Grabar and Iris Eshkol. 2016. Why do we reformulate? Automatic Prediction of Pragmatic Functions. In *HrTAL*, Dubrovnik, Croatia.
- Martine Grice and Stefan Baumann. 2007. An introduction to intonation functions and models. In Jürgen Trouvain and Ulrike Gut, editors, *Non-Native Prosody*, Trends in Linguistics. Studies and Monographs [TiLSM], pages 25–51. De Gruyter, Berlin and New York.
- Jonathan Hamaker, Yu Zeng, and Joseph Picone. 1998. Rules and guidelines for transcription and segmentation of the Switchboard large vocabulary conversational speech recognition corpus. Technical Report Version 7.1, Institute for Signal and Information Processing, Mississippi State University.
- Karin Pittner and Judith Berman. 2013. *Deutsche Syntax: Ein Arbeitsbuch*. Narr Studienbücher. Narr, Tübingen, 5 edition.
- Jochen Rehbein, Thomas Schmidt, Bernd Meyer, Franziska Watzke, and Annette Herkenrath. 2004. Handbuch für das computergestützte Transkribieren nach HIAT. Arbeiten zur Mehrsprachigkeit, 56.
- Marc Reznicek, Anke Lüdeling, Cedric Krummes, Franziska Schwantuschke, Maik Walter, Karin Schmidt, Hagen Hirschmann, and Torsten Andreas. 2012. Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01.
- Thomas Schmidt and Wilfried Schütte. 2010. FOLKER: An annotation tool for efficient transcription of natural, multi-party interaction. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 2091–2096.
- Thomas Schmidt and Swantje Westpfahl. A study on gaps and syntactic boundaries in spoken interaction. (submitted).
- Thomas Schmidt, Wilfried Schütte, and Jenny Winterscheid. 2015. cGAT. Konventionen für das computergestützte Transkribieren in Anlehnung an das Gesprächsanalytische Transkriptionssystem 2 (GAT2).
- Thomas Schmidt. 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*), pages 236–240.
- Thomas Schmidt. 2014a. The database for spoken german dgd2. In *Proceedings of the Ninth conference on International Language Resources and Evaluation (LREC'14)*, pages 1451–1457.
- Thomas Schmidt. 2014b. The research and teaching corpus of spoken german folk. In *Proceedings of the Ninth* conference on International Language Resources and Evaluation (LREC'14), pages 383–387.

- Margret Selting, Peter Auer, Dagmar Barth-Weingarten, Jörg Bergmann, Pia Bergmann, Karin Birkner, Elizabeth Couper-Kuhlen, Arnulf Deppermann, Peter Gilles, Susanne Günthner, Martin Hartung, Friederike Kern, Christine Mertzlufft, Christian Meyer, Miriam Morek, Frank Oberzaucher, Jörg Peters, Uta Quasthoff, Wilfried Schütte, Anja Stukenbrock, and Susanne Uhmann. 2009. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). Gesprächsforschung, 10:353–402.
- Rosmary Stegmann, Heike Telljohann, and Erhard W Hinrichs. 2000. Stylebook for the German Treebank in VERBMOBIL. Technical report, Technical Report 239, Verbmobil.
- Heike Telljohann, Erhard W Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2006. Stylebook for the Tübingen treebank of written German (TüBa-D/Z). In *Seminar fur Sprachwissenschaft, Universitat Tubingen, Tubingen, Germany*.
- Zahar Weiß and Gohar Schnelle. 2016. Frühneuhochdeutsche Satzsegmentierung: Annotationsrichtlinien: Version 4.0.
- Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006).*
- Angelika Wöllstein. 2010. Topologisches Satzmodell, volume 8 of Kurze Einführungen in die germanistische Linguistik. Winter, Heidelberg.
- Angelika Wöllstein. 2014. Topologisches Satzmodell. In Jörg Hagemann, editor, *Syntaxtheorien: Analysen im Vergleich*, Stauffenburg-Einführungen, pages 143–164. Stauffenburg-Verlag, Tübingen.