A Study on Gaps and Syntactic Boundaries in Spoken Interaction

Thomas Schmidt
Institut für Deutsche Sprache
R5, 6-13
D-68161 Mannheim
thomas.schmidt@ids-mannheim.de

Swantje Westpfahl
Institut für Deutsche Sprache
R5, 6-13
D-68161 Mannheim
westpfahl@ids-mannheim.de

Abstract

We present a study on gaps in spoken language interaction as a potential candidate for syntactic boundaries. On the basis of an online annotation experiment, we can show that there is an effect of gap duration and gap type on its likelihood of being a syntactic boundary. We discuss the potential of these findings for an automation of the segmentation process.

1 Introduction

1.1 Segmentation and the SegCor Project

The question of how to segment natural spoken interactions into units with a status equivalent to the sentence in written language is of importance not only from a theoretical point of view, but also with respect to research practices in corpus linguistics and the application of Computational Linguistics or Natural Language Processing tools (e.g. parsing) to transcripts of spoken language.

A great variety of segmentation principles for oral language have been proposed since the beginning of research on talk-in-interaction. However, we still lack a segmentation system that is both theoretically well-founded and practically operationalizable for large and diverse corpora of spoken interaction, and this impairs the use of such corpora for linguistic analysis, for language teaching, for contrastive studies and for the development of language technology.

The SegCor project has therefore set itself the aim to develop a method of segmentation that is adequate for the analysis of data from talk-in-interaction at different levels (such as syntax or interactional units) and for various communities of researchers. It evaluates and further develops approaches to segmentation put forward in the literature on conversation analysis, interactional linguistics, pragmatics and corpus linguistics by applying them to samples from three large collections of French and German audio and video recordings of various interaction types (the databases CLAPI, ESLO and FOLK, respectively). The project ultimately aims at a systematic segmentation guideline applicable across different interaction types and to French as well as German data.

Methodologically, the project approaches its subject matter from two different perspectives: 1) a qualitative, multidimensional approach which considers segmentation indices, problems and criteria and leads to tested and improved segmentation guidelines and 2) a quantitative, unidimensional approach based on selected criteria where possible boundaries are automatically identified and classified by human annotators according to their relevance for segmentation. Ideally, this second perspective will uncover a concrete potential for automatizing parts of the segmentation task. In this paper, we present a study carried out in the second perspective, restricted to the German data.

1.2 Syntactic Segmentation Guidelines

Our work in the project so far, which we cannot discuss in detail here, clearly indicates that a segmentation based primarily on syntactic criteria is, in general, superior in terms of robustness, intersubjectivity and practical applicability when compared to approaches based primarily on prosodic (e.g. Selting et al. 2009) or pragmatic (e.g. Rehbein et al. 2004) properties of speech.

Therefore, we have developed a segmentation guideline based on Topological Field Theory (cf. e.g. Pittner and Berman 2013 or Wöllstein 2014) as a model for identifying syntactic units for German data. After several iterations through a cycle of guideline refinement and evaluation including tests on inter-rater agreement, we have now reached a state where we can consider these guidelines as stable, and the resulting annotations as sufficiently reliable (see Westpfahl and Gorisch 2018). On the highest structural level, the guidelines define "Maximal Units (MUs)" as the fundamental unit of segmentation. We distinguish four different segment types: simple and complex sentential units (corresponding roughly to simple and complex sentences in written language), non-sentential units (such as backchannels) and abandoned (i.e. syntactically and/or pragmatically incomplete) units. Section 2.3. explains this in more detail. We use these units as the basis for the experiment described in what follows.

1.3 Gaps as Candidates for Syntactic Boundaries

Potentially, the end of every word (whether completed or not) in spoken language is a candidate for a segment boundary, but not all candidates are equally likely to actually *be* boundaries. Intuitively, words immediately followed by a gap – that is, an interval where the speaker's speech is interrupted for a short, but noticeable amount of time –, have an increased likelihood of constituting a segment boundary. Gaps can either be pauses (that is, "empty" silences), or they can be "filled" by another speaker's speech.[1]

In the FOLK corpus (Schmidt 2017), all pauses of at least 200ms duration are transcribed by trained transcribers according to cGAT (Schmidt et al. 20??). The current version of the corpus data uses the resulting inter-pausal units as one criterion (the other being speaker change, see below) to define a "contribution" as the fundamental segment in the data structure. In the absence of reliable alternatives, we chose this method of initial segmentation because it is theory-agnostic and makes segmentation a largely "mechanic" (i.e. objective, non-interpretative) decision for the transcribers. We now have a 2.25 million token corpus with this kind of initial segmentation. The corpus is fully time-aligned with the underlying audio and/or video recordings; each transcribed token is mapped to its standard orthographic equivalent, and annotated with lemma and POS information (see Westpfahl/Schmidt 2016). Visualization and query of those data could be much improved by switching to a more theory-grounded segmentation system. Being able to (partly) automate that task would reduce manual annotation labor[2] and allow us to continue keeping the initial transcription free from theory-dependent decisions.

As Example (1) shows, the end of a contribution (0047) can coincide with the end of a syntactic unit. In these cases, the intervening pause (line 0048) indicates a segment boundary.

Example (1)[3]:

0047 PB: °h (.) Flugtickets ham wir keine.
°h (.) We don’t have plane tickets.
0048 **(0.46)**
0049 PB: Foto nimmst du mit.
You bring the camera.

In other cases, such as Example (2), however, a syntactic unit is distributed across two contributions (0961 and 0963), and the intervening pause does *not* constitute a segment boundary.

Example (2):

0961 HK: Wie verhält sich
How behaves
0962 **(0.22)**
0963 HK: Josef K.?
Josef K.?

Example (3):

0007 DL: […] ob (.) die schon hier (unter) äh hinterlassen worden is.
[…] whether it has been deposited here already.
0008 CH: **Hast du deinen Studierendenaus[weis dabei]?**
Do you have your student ID with you?
0009 DL: [Ja, hab ich], eine Sekunde.
Yes, I do, just a second.

Example (3) also qualifies as a gap. Between contributions 0007 and 0009, speaker DL's speech is interrupted for about 1.1 seconds. In this case, however, instead of a silence, the gap is

[1] Other plausible candidates for segment boundaries could be filled pauses (i.e. hesitation markers like “äh”) or repair sequences. We are not considering those in the present study, but similar experiments could be carried out to analyze the boundary status of such units.

[2] The total number of gaps in the corpus of the kind which is discussed in this paper amounts to 177134.

[3] Punctuation and capitalization are added here for the sake of readability. The original transcripts follow the cGAT conventions for minimal transcripts and do not contain punctuation or capitalization. The following special symbols are used:

- °h, °hh, h° hh° represents audible breathing
- (.) represents a micropause, i.e. a noticeable silence shorter than 0.2s
- (0.46) represents a silence of 460ms
- square brackets represent overlapping speech

Registered DGD users can access the example at the URL given in the references section.

filled by speaker CH's question 0008 (whose last part overlaps with the start of 0009). The duration of this gap can be determined through the time-alignment of the contributions.

As Example (4) illustrates, such filled gaps (0458) can also lead to a syntactic unit being distributed over two contributions (0457 and 0459).

Example (4):

0457 TN:	aber die Selbständigkeit würd ich an ihrer Stelle auch mit im Blick behalten (.) weil ich *But, in your place, I would bear in mind self-employment as well (.) because I*
0458 DO:	**ab[solut]** *absolutely*
0459 TN:	[glaub sie s]ind ds (.) so vom Typ her […] *think you are like that (.) the type of person you are […]*

The study presented here aims at finding statistical evidence for a correlation between certain properties of such gaps and whether they lie *between* two syntactic segments (as in Examples 1 and 3) or *within* a syntactic segment (as in Examples 2 and 4).
We want to test the following hypotheses:

- The length of a gap can indicate whether there is a syntactic boundary or not,
- the type of a gap can indicate whether there is a syntactic boundary or not, and
- the parts of speech surrounding the gap can indicate whether there is a syntactic boundary or not.

Analyzing the results, we hope to distill some factors for an automatized segmentation process of our data.

1.4 Related Work

To our knowledge, research on the correlation between speaker pauses and syntactic segment boundaries is scarce or even non-existing for German spoken language interactions.

Yang (2007) presents a study comparing English publicly broadcasted language with Mandarin private conversations analyzing in how far speaker pauses mark the boundaries between "minor phrases [which] are clauses and phrases like PP, NP, VP, and fragments." (Yang 2007: 458). She analyzes whether there are dependencies between the length of the pause and their function as a boundary and points out that this depends on the "degree of spontaneity, as well as cognitive and communicative effort in conversational speech" (Yang 2007: 461).

Most studies on pauses are conducted in the context of psycholinguistic, sociolinguistic and pragmatic research. Psycholinguistic studies look at the cognitive reasons for speakers to pause their speech (cf. Beattie and Shovelton (2002) and Jong (2016)). These studies show that pauses between sentences are used mainly for conceptual planning, while pauses within sentences significantly correlate with less frequent word forms, indicating difficulties with lexical retrieval (Jong 2016).

With respect to sociolinguistic factors, Kendall (2009) could show that pause and speech rate vary by region, ethnicity, and gender.

Pragmatic studies focus on the ways pauses can fulfil "distinctive non-segmenting communicative function[s]" (Mukherjee 2001), i.e. functions other than segmenting or signaling hesitation. (cf. also Chafe 1995 and Rühlemann et al. 2011)

All of these studies rely on data which has already been segmented manually, and some of them explicitly exclude non-sentence-like data from their studies. Moreover, most of these studies focus on a specific interaction type, namely narratives.

Research on automatic segmentation of spoken language is either focused on the correction of segments in machine translation (Paulik et al. 2008), on evaluating machine translation output with possibly erroneous sentence boundaries (Matusov et al. 2005), on the automatic conversion of speech to text (Kolář 2008), or on speech synthesis (Holsteijn 1993 or Kock 2007).

Automatic segmentation systems as described in Kolář (2008) or also in the VERBMOBIL project (Kohler 1995 as described in Gibbon et al. 1997) are based on the prosodic analysis of the speech signal. In our FOLK corpus, we gather our data in everyday social interactions. This implies that few of our recordings have laboratory quality. Moreover, a lot of our interactions contain simultaneous contributions of two or more speakers. Tools based on the automatic processing of the audio signal do not work on our data. In addition, apart from the VERBMOBIL project, none of the studies were conducted on German.

With our study, we aim at shifting the perspective by first looking at the pauses (or gaps) in order to identify their potential for syntactic segmentation.

2 Experiment Setup

2.1 Sampling

The sample used for our experiment is drawn from altogether 259 interactions in version 2.8 of the FOLK corpus which cover a large variety of interaction types. In a first step, we randomly selected 200 of the 259 interactions and from each extracted, randomly again, 5 pairs of contributions C1 and C2 with the following properties:

- C1 and C2 have the same speaker.
- The time interval (i.e. the "gap") between the end of C1 and the start of C2 is at most 2.0 seconds long.
- Neither C1 nor C2 contain speech which was marked as incomprehensible by the transcriber.

This results in a random sample of 1.000 such pairs of contributions.

2.2 Classification of Gaps

We classify gaps according to two criteria. The first criterion is their duration, i.e. the length of the interval between the end of the first and the start of the second contribution. Given that the sampling already restricts gaps to a maximum length of 2.0 seconds, we chose to work with a division into four classes as shown in Table 1.

The second criterion is the gap type as illustrated above: we differentiate between gaps which are silences (as in Examples 1 and 2) and gaps which are not (as in Examples 3 and 4). As Table 1 shows, the data from the sample are not equally distributed across the resulting 4x2 matrix of duration/type combinations. Clearly, shorter gaps are more frequent than longer ones, and silence gaps are less frequent than their counterparts for all durations above 0.5 seconds. Since the sample is a random one, we can assume that it is representative of the entire corpus in this respect.

Duration d	Silence	Other	Total
d ≤ 0.5s	273	176	449
0.5s < d ≤ 1.0s	123	151	274
1.0s < d ≤ 1.5s	54	127	181
1.5s < d ≤ 2.0s	18	78	96
Total	468	532	1000

Table 1: Gaps in the sample according to duration and type

2.3 Annotation Guidelines

For the annotation experiment, we distilled a simplified version from the segmentation and annotation guidelines written for the SegCor project. The main focus of the guidelines for this experiment is to identify syntactic dependencies and to decide whether the gaps between two speaker contributions occur within dependent structures or mark the boundary between structurally complete units.

We use four annotation values with decreasing "boundariness" and an additional category for undecidable cases:

	Annotation value
1	Gap between MUs
2a	Gap between clauses within a MU
2b	Gap within a clause (within a MU)
2c	Gap within a word
3	Undecidable

Table 2: Classification options for the annotators

The smallest unit is a word and thus the annotator has to decide whether the gap is within a word (annotation value = 2c) as in the following example in which the word “versetzen” (relocate) is distributed over two contributions, i.e. there is a gap between the prefix “ver-“ (re-), which cannot stand alone, and the stem “setzen” (locate):

Example (5):

0628 JA:	Dann hast du ja auch keine mehr zu ver *Then you don’t have any left to re-*
0629	**(0.41)**
0630 PA:	[fünf, se]chs, sieben, genau. *five, six, seven, exactly.*
0631 JA:	[setzen]. *-locate.*

The second smallest syntactic unit is a clause. Clauses can be either main clauses, i.e. in German with the finite verb in the first or second position, or subordinate clauses, i.e. in German with the finite verb in the last position such as relative clauses etc. The annotator has to decide whether the gap in the contribution is within a clause (annotation value = 2b) as in Example 2 above. Here, the subject of the main clause, "Josef K.", is only uttered after a gap of 0.22 seconds.

With the identification of clauses it is possible to identify larger, complex syntactic units and whether a gap is situated between two clauses but within a complex syntactic unit, i.e. between a main and a subordinate, relative, infinitive, or conditional clause, or between two coordinated main clauses if and only if one of the two clauses shows subject or verb ellipsis.

The following example illustrates a gap between a main and a subordinate clause which would be annotated with the value 2a:

Example (6):

0090 EG: ähm ja. Die sind jetz aber nich gekommen,
uhm yes. They didn't come
0091 **(0.48)**
0092 EG: äh weil (.) der meinte dann gleich so: „ja […]“
uh because (.) he directly said: “yes […]”

Finally, a maximal (syntactic) unit can consist of

- compound sentences as in Example 6,
- simple sentences such as main clauses without subordinate clauses,
- sentence equivalents (i.e. non-sentential units), and
- disrupted utterances (i.e. abandoned units).

Sentence equivalents are defined on a pragmatic level as they only show limited syntactic dependencies (e.g. nominal phrases). They cannot be defined syntactically because they lack a finite verb. However, pragmatically, they can be considered complete, e.g. nominal phrases, prepositional phrases etc., or speech particles such as interjections, responsives, backchannel and reception signals, vocatives etc.

Example (7):

0078 JS: °h genau ähm h°
°h exactly uhm h°
0079 **(0.31)**
0080 JS: Jetzt ham sie sich mit Versprechern beschäftigt.
Now you have occupied yourself with slips of the tongue.

The gap of 0.31 seconds in line 0079 of Example 7 is considered as a gap between two maximal units, i.e. a non-sentential unit “genau” (exactly) and a simple sentence. Gaps like this will be annotated with the value (1) in our annotation experiment and will be interpreted as boundaries of segments.

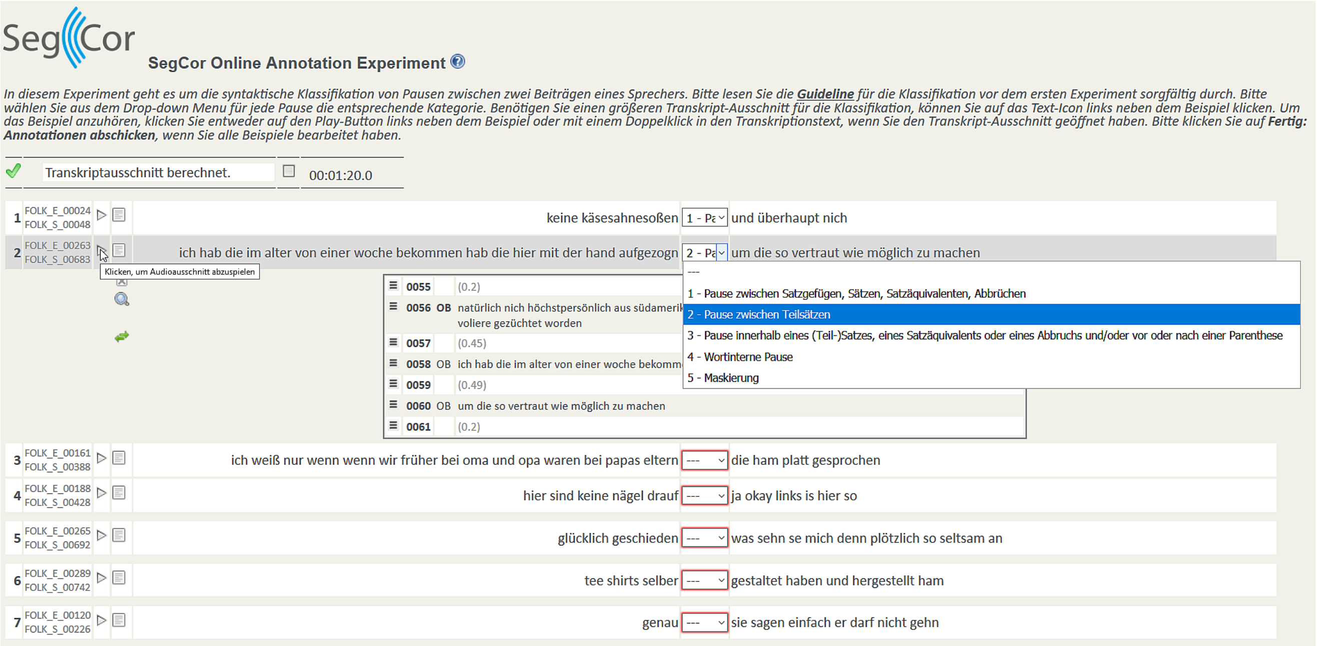
In this scheme, also abandoned units are considered as maximal units. They are defined by utterances opening up a syntactic projection which is not fulfilled in the following utterances.

Examples 6 and 7 also illustrate a specific rule of the guidelines with respect to the interpretation of phenomena typical for transcripts of spoken language such as transcribed audible breathing (°h, h°) and hesitation particles (“äh”, “ähm”). They are not considered as units on their own but rather part of the preceding or following segment except when they are surrounded by gaps.

Finally, we provide a category for cases which are undecidable (annotation value = 3), i.e. ambiguous in their syntactic structure. This is the case, e.g. when their interpretation strongly relies on the prosody yet the audio is masked because of a mentioned name etc.

2.4 Online Annotation Environment

We chose to implement the annotation task in a web-based environment which is connected to the architecture of the Database for Spoken German (DGD, see Schmidt 2017). The extra effort this entails is justified by practical considerations: first, by putting the experiment online in this way we profit from the possibility of integrating existing DGD functionality (such as audio playback) into the annotation GUI. Second, it makes the annotation task accessible to all 8.000 registered users of the DGD, so we have the option to "crowd-source" annotations. In each round of annotation, annotators are presented 15 randomly chosen pairs from the sample in a KWIC-like format as shown in Figure 1. Note that no written information on gap duration or type is given. If the contributions preceding and following the gap are not sufficient to decide, users can extend the context to further contributions and they can playback the corresponding audio in case the transcription itself does not provide enough information. Once all 15 gaps have been annotated, i.e. assigned one of the values listed in Table 2, the result is sent to the server. Annotators can then choose to terminate the experiment or to do another round of annotations. The experiment is online and open to all registered DGD users at the URL given in the references section.



SegCor Online Annotation Experiment

*In diesem Experiment geht es um die syntaktische Klassifikation von Pausen zwischen zwei Beiträgen eines Sprechers. Bitte lesen Sie die **Guideline** für die Klassifikation vor dem ersten Experiment sorgfältig durch. Bitte wählen Sie aus dem Drop-down Menu für jede Pause die entsprechende Kategorie. Benötigen Sie einen größeren Transkript-Ausschnitt für die Klassifikation, können Sie auf das Text-Icon links neben dem Beispiel klicken. Um das Beispiel anzuhören, klicken Sie entweder auf den Play-Button links neben dem Beispiel oder mit einem Doppelklick in den Transkriptionstext, wenn Sie den Transkript-Ausschnitt geöffnet haben. Bitte klicken Sie auf **Fertig: Annotationen abschicken**, wenn Sie alle Beispiele bearbeitet haben.*

✓ Transkriptausschnitt berechnet. □ 00:01:20.0

1	FOLK_E_00024 FOLK_S_00048	keine käsesahnesoßen	1 - Pa	und überhaupt nich
2	FOLK_E_00263 FOLK_S_00683	ich hab die im alter von einer woche bekommen hab die hier mit der hand aufgezogn	2 - Pa	um die so vertraut wie möglich zu machen
3	FOLK_E_00161 FOLK_S_00388	ich weiß nur wenn wenn wir früher bei oma und opa waren bei papas eltern	---	die ham platt gesprochen
4	FOLK_E_00188 FOLK_S_00428	hier sind keine nägel drauf	---	ja okay links is hier so
5	FOLK_E_00265 FOLK_S_00692	glücklich geschieden	---	was sehn se mich denn plötzlich so seltsam an
6	FOLK_E_00289 FOLK_S_00742	tee shirts selber	---	gestaltet haben und hergestellt ham
7	FOLK_E_00120 FOLK_S_00226	genau	---	sie sagen einfach er darf nicht gehn

Klicken, um Audioausschnitt abzuspielen

0055		(0.2)
0056	OB	natürlich nich höchstpersönlich aus südamerik voliere gezüchtet worden
0057		(0.45)
0058	OB	ich hab die im alter von einer woche bekomm
0059		(0.49)
0060	OB	um die so vertraut wie möglich zu machen
0061		(0.2)

---
1 - Pause zwischen Satzgefügen, Sätzen, Satzäquivalenten, Abbrüchen
2 - Pause zwischen Teilsätzen
3 - Pause innerhalb eines (Teil-)Satzes, eines Satzäquivalents oder eines Abbruchs und/oder vor oder nach einer Parenthese
4 - Wortinterne Pause
5 - Maskierung

Figure 1: Screenshot of the online annotation experiment in the Database for spoken German (DGD). Similarly as in the DGD, one can extend the context of the contribution pair and listen to the corresponding audio file. One can choose between five categories, 2-4 corresponding to the presented annotation values of 2a, b, and c, and 5 corresponding to the annotation value 3, presented in Table 2.

3 Results

3.1 Overview

In our pilot experiment, 11 annotators (recruited among project members and FOLK student assistants) carried out 100 rounds of 15 annotations, resulting in altogether 1500 annotated contribution pairs. Owing to the random choice of pairs for each round, some (more precisely: 232 of 1000) pairs from the sample remained unannotated, while others were annotated more than once (by one and the same or by different annotators).[4] All following calculations are based on all annotations, i.e. double annotations (whether conflicting or agreeing) were not treated in any special way. A simple count of all 1500 annotation values reveals that a majority of gaps were classified as coinciding with an MU boundary (Table 3).

	Annotation value		
1	Between MUs	927 (62%)	
2a	Between clauses	104 (7%)	481 (32%)
2b	Within a clause	362 (24%)	
2c	Within a word	15 (1%)	
3	Undecidable	92 (6%)	

Table 3: Annotation values by category

This gives us a baseline for automatic segmentation: if we simply classified all gaps as MU boundaries (as is the de facto state in FOLK now), we would be wrong in at most 38% of all cases. In what follows, we will now analyze if and how the distribution changes when we take into account the duration and type of gaps.

3.2 Gap Duration

In order to measure dependency on gap duration, we conflate annotation values 2a, b and c into one category and ignore cases which annotators marked as undecidable. We thus keep a binary distinction between "boundary" (1) and "no boundary" (2a-c).

The numbers in Table 4 confirm what is intuitively plausible: longer gaps are more likely to coincide with an MU boundary than shorter ones, or – taking the perspective of the speaker: "You don‘t pause for too long when you haven‘t finished yet". Compared to the overall mean of 62%, gaps up to 0.5s have a decreased, all other duration classes an increased likelihood of being a boundary. The longer the gap, the more likely it gets that the gap is a boundary (up to 91% for gaps between 1.5 and 2 seconds).

	d ≤ 0.5	ALL	0.5<d≤ 1	1<d≤ 1.5	1.5<d≤ 2
1	385 (50%)	927 (62%)	250 (70%)	183 (75%)	109 (91%)
2a-c	344 (44%)	481 (32%)	85 (24%)	44 (18%)	8 (7%)

Table 4: Dependency on gap duration (percentages are the relative frequencies of the respective annotation values for the different gap durations)

3.3 Gap Type

We use the same calculation method for looking at dependency on gap type. Again, there is an obvious tendency: a filled gap is more likely to occur at a segment boundary than a simple silence. Intuitively, this is best explained by taking the perspective of the speaker who fills the gap: "If the other speaker is in the middle of some construction, there is a reduced tendency to take the turn or provide a backchannel."

	Silence	ALL	Other
1	408 (53%)	927 (62%)	519 (71%)
2a-c	320 (41%)	481 (32%)	161 (22%)

Table 5: Dependency on gap type (percentages are the relative frequencies of the respective annotation values for the different gap types)

3.4 Gap Duration and Type

If we combine the two types of gap classification, we end up with the following matrix showing the likelihood of being a boundary for a given duration/type combination:

Duration d	Silence	Other
d ≤ 0.5s	46%	63%
0.5s < d ≤ 1.0s	70%	79%
1.0s < d ≤ 1.5s	75%	84%
1.5s < d ≤ 2.0s	76%	98%

Table 6: Combined dependency on gap duration and type

While the figures clearly support the hypothesis that there is a correlation between gap duration/type and “boundariness”, they also show that these parameters are not sufficient as a basis for an automatic classification. If we now "close the gap", i.e. merge the respective contributions, for the one type of combination which has less

[4] Of the 768 remaining annotated gaps, 426 gaps were annotated at least twice, either by the same or by different annotators. The inter-rater-agreement turned out to be similar to the results of our previous annotation guidelines tests (a Kohen’s kappa of .69, Westpfahl/Gorisch 2018).

than half a chance of constituting a boundary (silence with d ≤ 0.5s) and assign boundary status to all the other combinations, we obtain an overall error rate of 30%[5]. This is only a minor improvement over the baseline, which is explained by the fact that combinations with a clearer tendency are rarer in the data, or, conversely, that the tendency is largely unclear (i.e. close to 50%) for the most frequent combination. An error rate of 30% is a long way off an acceptable precision, both for a fully automated task and as a basis to speed up manual segmentation. We will therefore need to consider additional parameters with an effect on the likelihood for a gap to be a boundary.

3.5 Part-Of-Speech

It is a plausible hypothesis that the likelihood for a syntactic boundary between two tokens A and B is sensitive to A's and B's parts of speech. For instance, one would expect fewer boundaries to occur between a pronoun and a verb (as in Example 4 above) than between a verb and a response particle (Example 3). The same can be assumed when A and B are additionally separated by a gap. Hence, we looked for tendencies in our data in this respect. To this end, we first classified each gap according to the POS combinations of the tokens which precede and follow it. The data were originally tagged with the TreeTagger using a parameter file for spoken German and the STTS 2.0 (Westpfahl et al. 2017). We know that this procedure attains a precision of around 95% on FOLK. Since the tagset is rather detailed, we reduced the number of categories by mapping each POS (e.g. VVFIN) to one of 10 superordinate categories (e.g. V). We then discarded all combinations that occurred less than five times. From the remaining 49 combinations, some show a clear tendency to increase or decrease boundary likelihood. Table 7 lists some interesting examples with such a tendency.

For example, the combination of two adjectives or adverbs (A-A), a sequence of a noun and an article (N-ART) or a sequence of an adjective and a verb (A-V) all coincide relatively rarely with a boundary, whereas boundary likelihood is increased for sequences of a noun and a pronoun (N-P). It is also noticeable that all other combinations with increased boundary likelihood contain in one position a token which STTS 2.0 classifies as "non-grammatical" (NG, e.g. responsives, interjections or hesitation markers) or "sentence-external" (SE, e.g. discourse markers such as “also”), so these POS may generally be good indicators for boundaries.

POS combinations are thus a candidate for an additional parameter for identifying boundaries (of course not just around gaps, but potentially also at other positions). However, given that there are many more POS combinations than gap types or durations, we feel that absolute numbers from the first round of our experiment are not sufficiently high to derive reliable statistics from them.

POS	**1**	**2a-c**	**Boundary?**
N-PTK	0	26	0%
N-ART	1	10	9.1%
A-A	3	25	10.7%
V-AP	4	21	16.0%
KO-P	2	9	18.2%
A-V	6	17	26.1%
NG-A	29	4	87.9%
NG-P	64	8	88.9%
NG-NG	98	12	89.1%
P-NG	26	3	89.7%
N-P	38	3	92.7%
NG-KO	17	1	94.4%
PTK-NG	21	1	95.5%
V-NG	69	2	97.2%
A-NG	36	1	97.3%
A-SE	10	0	100%

Table 7: Dependency on POS (selection)

4 Conclusion and Outlook

We have shown that the duration and type of gap between two contributions have an effect on their likelihood of being a syntactic boundary. However, although the tendencies are clear, these parameters alone are not sufficient as a basis for an automatic segmentation process. Additional parameters will have to be evaluated and integrated into the statistics before we can hope to obtain an acceptable precision. As we have shown, POS combinations are one candidate for such a parameter. Others may be the relative frequency of the word following the gap (see Jong 2016 in

[5] This error rate is calculated as follows: for silence with d ≤ 0.5s, 46% of 273 instances (see table 1) would be classified incorrectly; for silence with 0.5s < d ≤ 1s, 30% (=1.0 - 0.7) of 123 instances would be classified incorrectly, and so forth. In total, 300 of 1000 instances (=30%) would be classified incorrectly.

related work) or general characteristics of the interaction, of speakers or of the respective contributions (e.g. interactions type, speech rate, overall distribution of gaps). We will explore these parameters in future work of the project. Ideally, a combination of suitable parameters in a multi-factorial model will make possible a fully automatic decision on the boundary status of gaps. More modestly, we hope at least to be able to make a sufficiently reliable decision in a sufficiently large number of cases so that we can reduce manual annotation effort by restricting it to those cases where the statistics do not allow a clear classification.

The online experiment has proven to be an adequate means of obtaining larger numbers of annotations on which we can base our statistics. We will attempt to broaden the audience in the next phase, possibly slightly simplifying the experiment setup on that occasion (e.g. by dispensing with double annotations). If we manage to increase the number of annotations by one order of magnitude (i.e. by targeting 15,000 annotated contribution pairs), the numbers on POS combinations should become reliable enough to be integrated into the statistics. Similar experiments, for instance on syntactic boundaries in the neighborhood of hesitation markers, can be carried out to gain insight into statistics of syntactic boundaries in other positions.

Acknowledgments

The work presented here is financed by a joint grant from the German Research Foundation (DFG) and the French National Research Agency (ANR).

References

Geoffrey Beattie and Heather Shovelton. 2002. Lexical access in talk. A critical consideration of transitional probability and word frequency as possible determinants of pauses in spontaneous speech. *Semiotica,* 141:49–71.

Wallace L. Chafe. 1995. Some reasons for hesitating. In Deborah Tannen, editor, *Perspectives on silence*. 2nd print. Ablex, Norwood, NJ:77–92.

Dafydd Gibbon, Roger Moore and Richard Winski. 1997. *Handbook of standards and resources for spoken language systems.* Mouton de Gruyter, Berlin and New York.

Nivja H. de Jong. 2016. Predicting pauses in L1 and L2 speech. The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching,* 54(2):24.

Tyler S. Kendall. 2009. *Speech rate, pause, and linguistic variation: An examination through the Sociolinguistic Archive and Analysis Project.* Dissertation, Duke University, Durham, North Carolina.

Adrien S. Kock. 2007. *Enhancing Synthetic Speech with Filled Pauses.* Habilitation, University of Twente, Department of Computer Science chair of Human Media Interaction. Enschede, Netherlands.

Klaus J. Kohler. 1995. *From scenario to segment: The controlled elicitation, transcription, segmentation and labelling of spontaneous speech (Arbeitsberichte / Institut für Phonetik und digitale Sprachverarbeitung).* University of Kiel (Institut für Phonetik und digitale Sprachverarbeitung), Germany.

Jáchym Kolář. 2008. *Automatic Segmentation of Speech into Sentence-like Units.* Dissertation, University of West Bohemia, Pilsen, Czech Republic.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating Machine Translation Output with Automatic Sentence Segmentation. *Proceedings of the International Workshop on Spoken Language Translation (IWSLT).* Pittsburgh, PA, USA.

Joybrato Mukherjee. 2001. Speech is Silver, but Silence is Golden. Some Remarks on the Function(s) of Pauses. *Anglia - Zeitschrift für englische Philologie,* 118(4):571–584.

Matthias Paulik, Sharath Rao, Ian Lane, Stephan Vogel, and Tanja Schultz. 2008. Sentence segmentation and punctuation recovery for spoken language translation. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing.* IEEE, Las Vegas, NV, USA.

Karin Pittner and Judith Berman. 2013. *Deutsche Syntax. Ein Arbeitsbuch.* 5th edition. Narr, Tübingen, Germany.

Jochen Rehbein, Thomas Schmidt, Bernd Meyer, Franziska Watzke, and Annette Herkenrath. 2004. Handbuch für das computergestützte Transkribieren nach HIAT. *Arbeiten zur Mehrsprachigkeit* (56).

Christoph Rühlemann, Andrej Bagoutdinov, and Matthew Brook O'Donnell. (2011): Windows on the Mind. Pauses in Conversational Narrative. *IJCL*, 16(2):198–230.

Margret Selting, Peter Auer, Dagmar Barth-Weingarten, Jörg Bergmann, Pia Bergmann, Karin Birkner, et al. 2009. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung*, 10:353–402.

Thomas Schmidt, Wilfried Schütte, and Jenny Winterscheid. 2015. *cGAT. Konventionen für das computergestützte Transkribieren in Anlehnung an das Gesprächsanalytische Transkriptionssystem 2 (GAT2). Version 1.0, November 2015.* IDS Mannheim, Germany.

Thomas Schmidt. 2017. Construction and Dissemination of a Corpus of Spoken Interaction – Tools and Workflows in the FOLK project. *Corpus Linguistic Software Tools, Journal for Language Technology and Computational Linguistics (JLCL)*, 31(1):127-154.

Yvonne von Holsteijn. 1993. TextScan: A preprocessing module for automatic text-to-speech conversion. In Vincent van Heuven and Louis C. W. Pols, editors, *Analysis and Synthesis of Speech. Strategic Research towards High-Quality Text-To-Speech Generation*. Mouton de Gruyter (Speech research, 11), Berlin and New York:27–41.

Swantje Westpfahl and Thomas Schmidt. 2016. FOLK-Gold — A gold standard for part-of-speech-tagging of spoken German. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia*. European Language Resources Association (ELRA), Paris, France:1493–1499.

Swantje Westpfahl, Thomas Schmidt, Jasmin Jonietz, and Anton Borlinghaus. 2017. *STTS 2.0. Guidelines für die Annotation von POS - Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS). Version 1.1, March 2017*. IDS Mannheim, Germany.

Swantje Westpfahl and Jan Gorisch. 2018. A syntax-based scheme for the annotation and segmentation of German spoken language interactions. *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), Workshop at COLING 2018, Santa Fe, New Mexico*.

Angelika Wöllstein. 2014. Topologisches Satzmodell. In Jörg Hagemann, editor, *Syntaxtheorien: Analysen im Vergleich*. Stauffenburg-Verlag (Stauffenburg-Einführungen, 28), Tübingen, Germany:143–164.

Li-chiung Yang. 2007. Duration and pauses as boundary-markers in speech. A cross-linguistic study. *8th Annual Conference of the International Speech Communication Association. INTERSPEECH*. Antwerp, Belgium, August 27-31:458–461.

[Example 1]: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscript&id=FOLK_E_00030_SE_01_T_03_DF_01&cID=c47&wID=w139

[Example 2]: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscript&id=FOLK_E_00120_SE_01_T_01_DF_01&cID=c962&wID=w3360

[Example 3]: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscript&id=FOLK_E_00305_SE_01_T_01_DF_01&cID=c8&wID=w25

[Example 4]: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscript&id=FOLK_E_00174_SE_01_T_01_DF_01&cID=c457&wID=w1818

[Example 5]: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscript&id=FOLK_E_00132_SE_01_T_04_DF_01&cID=c628&wID=w2714

[Example 6]: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscript&id=FOLK_E_00084_SE_01_T_01_DF_01&cID=c91&wID=

[Example 7]: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscript&id=FOLK_E_00003_SE_01_T_01_DF_01&cID=c78&wID=w497

[Experiment-URL] http://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=contributionChainExperiment