Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5169–5177
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

GRAIN-S: Manually Annotated Syntax for German Interviews

Agnieszka Falenska[1], Zoltán Czesznak[1], Kerstin Jung[1], Moritz Völkel[1], Wolfgang Seeker[2], Jonas Kuhn[1]
[1] University of Stuttgart, Institute for Natural Language Processing (IMS)
[2] Retresco, Berlin
{firstname.lastname}@ims.uni-stuttgart.de, wolfgang.seeker@retresco.de

Abstract

We present GRAIN-S, a set of manually created syntactic annotations for radio interviews in German. The dataset extends an existing corpus GRAIN and comes with constituency and dependency trees for six interviews. The rare combination of gold- and silver-standard annotation layers coming from GRAIN with high-quality syntax trees can serve as a useful resource for speech- and text-based research. Moreover, since interviews can be put between carefully prepared speech and spontaneous conversational speech, they cover phenomena not seen in traditional newspaper-based treebanks. Therefore, GRAIN-S can contribute to research into techniques for model adaptation and for building more corpus-independent tools.

GRAIN-S follows TIGER, one of the established syntactic treebanks of German. We describe the annotation process and discuss decisions necessary to adapt the original TIGER guidelines to the interviews domain. Next, we give details on the conversion from TIGER-style trees to dependency trees. We provide data statistics and demonstrate differences between the new dataset and existing out-of-domain test sets annotated with TIGER syntactic structures. Finally, we provide baseline parsing results for further comparison.

Keywords: syntax, treebank, non-canonical data

1. Introduction

Treebanks, i.e. structurally annotated corpora, play an important role both in the language sciences (linguistics, psycholinguistics) and in speech and language technology. They serve as gold-standard data for testing hypotheses or evaluating automatic systems, provide the signal in supervised training of machine learning models, or inform processes of adaptation, generation of synthetic data, etc. Therefore, for more and more languages corpora annotated for syntactic structure have been provided to the research community – not least in response to the Universal Dependencies initiative (Nivre et al., 2016).

It is known that for language-technological systems trained with supervised machine learning, there is a relatively strong dependency on the text genre, language register, content domain and other dimensions of the material in the training corpus (Sekine, 1997). For research into techniques for model adaptation and for building more corpus-independent tools, it is important to have test data that represent relevant variations of existing treebanks for the same language. For example, since adaptation of text-processing tools to spoken language is of central importance to many research and application contexts, the availability of manually annotated syntactic structures on samples of spoken utterances is crucial.

In this contribution we present GRAIN-S(yntax) – a set of manually created syntactic annotations for GRAIN, a corpus of German RAdio INterviews (Schweitzer et al., 2018). The nature of the interview situation differentiates GRAIN-S from existing German treebanks. The utterances we find in the corpus can be put halfway between carefully prepared speech and spontaneous conversational speech: the interviewers' questions are presumably partially prepared, and in certain cases, the interviewees' answers reflect some previously thought through positions as well. In addition, many of the involved speakers can be considered as experienced speakers, working e.g. in political or public settings. This means that tool evaluation on the annotated data should not be taken to reflect average performance on spoken language analysis. Instead, the dataset can act as a "stepping stone" of semi-spontaneous spoken German for informed research into adaptation techniques, e.g., by drawing attention to the systematic differences between written and (non-read) spoken German that already come to the surface in partially planned utterances. Moreover, since the GRAIN interviews originally come with audio recordings and multiple layers of gold- and silver-standard annotations, extending them with syntactic structures creates a beneficial combination of text and speech annotations. Such combination can serve as a very valuable resource for multi-modal text- and speech-based research.

2. Related Work

Many long-lasting German treebanks are based on primary data from the news domain, such as TIGER (Brants et al., 2004), TüBa-D/Z (Hinrichs et al., 2004), or HDT (Foth et al., 2014). More specifically, TIGER and TüBa-D/Z contain German newspaper data and HDT online newscasts from a technical news service. More recent approaches, such as the Universal Dependencies Project (Nivre et al., 2016), introduce German treebanks containing articles from Wikipedia and historic literary text (see the latest release v2.5 of the Universal Dependencies (Zeman et al., 2019)).

NoSta-D (Dipper et al., 2013) and the test suite from Seeker and Kuhn (2014) provide common syntactic annotations for several domains. NoSta-D includes historical, chat and learner data, literary prose, newspaper texts and also spoken data from a map task. Seeker and Kuhn (2014) include DVD manuals, alpine hiking stories, text from a novel, proceedings from the European Parliament and economy news. Both datasets are based on the TIGER annotation.

	#sentences all/int./guest	#tokens	interviewer (gender)	guest (gender)
2014-05-24	94/29/65	1894	Rebecca Lüer (f)	Karl-Josef Laumann (m) *Pflege- und Patientenbeauftragter der Bundesregierung* State Secretary in the Federal Ministry of Health
2014-12-06	107/27/80	1954	Jan Seidel (m)	Michael Hüther (m) *Direktor des Instituts der deutschen Wirtschaft* Director of the German Economic Institute
2015-01-24	128/41/87	1848	Evelyn Seibert (f)	Rainer Wendt (m) *Bundesvorsitzender der Deutschen Polizeigewerkschaft* National Chair of the German Police Trade Union
2015-06-20	87/23/64	2025	Evi Seibert (f)	Holger Münch (m) *Präsident des BKA* President of the Federal Criminal Police Office
2015-08-08	109/32/77	1633	Rebecca Lüer (f)	Maria Krautzberger (f) *Präsidentin Umweltbundesamt* President of the German Environment Agency
2015-09-19	101/25/76	1920	Uwe Lueb (m)	Ingo Kramer (m) *Arbeitgeberpräsident (BDA)* President of the Confederation of German Employers' Associations

Table 1: GRAIN-S annotated interviews, total number of sentences: 626. Gender information is deduced from first names of the speakers.

While Seeker and Kuhn (2014) provide dependency trees from a conversion step (Seeker and Kuhn, 2012), NoSta-D is directly annotated with dependencies. Regarding further spoken primary data, the DIRNDL corpus (Eckart et al., 2012) comes with automatically annotated constituency trees based on the German LFG-grammar by Rohrer and Forst (2006). However, the primary data are also from the news domain (read radio news) and the syntactically sound manuscripts have been used for the syntactic annotation.
Nevertheless, approaches such as from Dannenberg et al. (2016) show, that there is an interest in syntactic analysis of spontaneous speech. They compare syntactic trees of American English data to the respective prosodic tree structures. However, they opt at a mostly automatic setting, thus also make use of automatically created syntactic analyses.

3. Characteristics of the Dataset

The original GRAIN corpus consists of 140 German radio interviews and comes with two parts: a silver-standard part consisting of over 10 layers of automatic speech and text annotations, and a gold-standard part, with 5 layers of manual annotations for a subset of 20 interviews. The gold-standard annotations of GRAIN are based on a textual version of the interviews which includes features of orality such as repetitions and broken syntax, but does not include partly uttered words or non-lexical fillers (such as "ähm" or "hm"). This is due to the fact, that GRAIN is based on two sets of primary data: the audio files of the broadcasts and a textual version, also provided by the radio station, which was highly edited for readability. The latter would not have posed a challenge to the text processing tools, and a close transcription of the audio files would have led to a huge case of unknown vocabulary. Thus, the gold-standard annotations were based on *unnormalizations* of the edited textual versions. A more detailed description of this process can be found in (Eckart and Gärtner, 2016).
Apart from textual unnormalization, the gold-standard part of GRAIN consists of manually annotated part-of-speech tags, referential information status (Riester and Baumann, 2017), questions-under-discussion (Reyle and Riester, 2016), and information structure (Riester et al., 2018). GRAIN-S expands this part by adding manually annotated syntactic trees for six of those interviews. Each of the interviews contains around 100 sentences which in total gives exactly 626 sentences and 11274 tokens. The interviews, as a part of SFB732 Silver Standard Collection, come with audio recording[1] and additional metadata, such as broadcast date, names of the speakers and their affiliation (see Table 1 for details).

4. Data Annotation

4.1. Part-of-speech Tags

The gold-standard part-of-speech annotation from GRAIN was used as a basis for the syntax annotation. The part-of-speech tags were considered as given, i.e. no changes to the part-of-speech layer were allowed during the syntax annotation to keep GRAIN-S compatible with the other manually created annotation layers.

4.2. Syntax Trees

The syntax annotation was done by two linguistically trained annotators in two rounds. In the first round, the annotators worked alone, then, in the second round they merged their results into one version.[2] Difficult cases were

[1] Available on request.

[2] We release results of both rounds, i.e., annotations of single annotators and the merged version.

Category/Label		Constituent	
S	SB	*Der*	Which
	AVP/MO	*auch heute*	also today
	S/MO	***, finde ich ,***	, I think ,
	NP/PD	*eine sehr anspruchsvolle Ausbildung*	a very high-quality education
	HD	*ist*	is
	PU	.	.

(a) Sentence 20140525.4: insertion annotated as a modifier.

Category/Label		Constituent	
S	JU	*Und*	And
	MO	*deswegen*	that's why
	S/PAR	***will ich jetzt nicht übertreiben ,***	I do not want to exaggerate now,
	HD	*reden*	are talking
	SB	*wir*	we
	MO	*eigentlich*	actually
	PP/OP	*von einem doch grundlegenden Umbau, auch der Philosophie der Pflegeversicherung*	about a fundamental change, even the philosophy of nursing insurance.
	PU	.	.

(b) Sentence 20140525.29: insertion annotated as parentheses.

Category/Label			Constituent	
S+/CC	PU		,	,
	S-/CJ		*dass er in seinem Unternehmen*	that he in his company
	PU		,	,
	S/CJ	**S/MO**	***wenn denn ein freier Arbeitsplatz da ist***	when there is an unoccupied position available
		PU	,	,
		CP	*dass*	that
		SB	*er*	he
		VP/OC	*den auch erfüllen*	take that position too
		HD	*kann*	can

(c) Part of the sentence 20150919.66: example usage of categories S+ and S-.

Figure 1: Examples of sentences with insertions (marked in bold). For readability reasons only annotations for top-level constituents are presented.

solved under supervision of an expert. To facilitate the attachment of tokens to their correct phrases, the annotators were allowed to listen to the original sound files of the interviews to gather intonational information, if necessary. The annotation tool *PhiTag*[3] was used for both creating and merging the annotations.

Regarding the guidelines of the annotation, we followed the annotation scheme of the TIGER corpus (Brants and Hansen, 2002).[4] The objective was to stay as close as possible to the original guidelines but in cases where domain-related phenomena were not captured by them. In such cases we had to adapt the framework to our task (compare also the set-up of the NoSta-D annotations spanning several non-standard varieties of German (Dipper et al., 2013)). TIGER contains newspaper articles which represent written and edited usage of language. By contrast, the interviews are samples of spoken conversations which yield sentences with many insertions, reparanda, and questions (see Section 6. for numeric differences between those two datasets). To annotate cases not covered by the original guidelines we introduced few changes described below.

Parentheses vs. modifiers. Both written and spoken sentences can contain insertions. TIGER guidelines define insertions as parts that carry extra information but cannot be syntactically integrated into the rest of the sentence. Such parts are annotated with the function label PAR (parenthesis). Since in speech similar constructs can be viewed as modifiers, we label them accordingly. To be more precise, if an insertion could be replaced with a simple phrase without semantic change, it is treated as a modifier and labeled with MO.

Figure 1 demonstrates examples of sentences with different types of insertions. In Figure 1a we find a modifying insertion, because *finde ich* (eng. I think) could be replaced by

[3] https://phitag.de

[4] Conversion from TIGER-style trees to Universal Dependencies is not straightforward and requires manual work. Therefore, we leave it for future work.

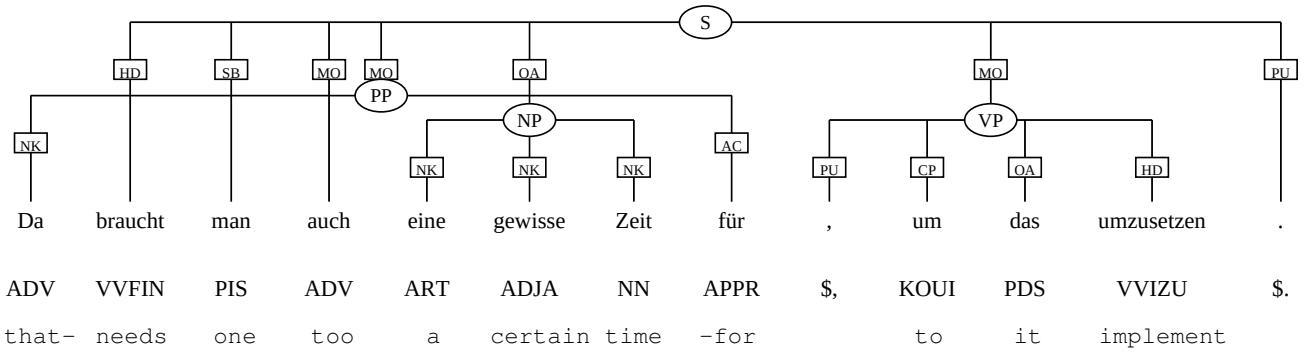


Figure 2: Sentence 20140524.35: example of preposition stranding.

the prepositional phrase (PP) *meiner Meinung nach* (eng. in my opinion). This PP integrates smoothly into the rest of the sentence. In contrast, the insertion in Figure 1b exemplifies a parenthesis. It is a clause, but it is neither coordinated, nor subordinated with respect to the rest of the sentence. Also, we cannot find a natural phrasal replacement for it.

Phrasal and incomplete sentences. Spontaneous speech can produce lower rank sentences. For example, (obvious) subjects can be dropped, even though German is not a pro-drop language. Discourses contain short phrasal utterances, without subject and/or predicate. There are reparanda in the corpus, i.e., sentences which are interrupted and then corrected by the speaker. Likewise, a sentence can be left unfinished because of an interrupting comment, and then get continued and/or slightly rephrased to match the introduced interruption.

As a consequence, in our annotation not all sentences have the category S as the root node. In case of one-phrase-sentences, we let phrasal nodes (e.g., noun phrases NP, adverbial phrases AVP) to be root nodes. For example, sentences 20141206.107: *Sehr gerne.* (eng. With pleasure.), 20150620.65: *Ja.* (eng. Yes.), or 20150919.95: *Teils, teils* (eng. Partly, partly.) were annotated as adverbial phrases ADV.

Moreover, we introduce a new category S- for incomplete sentences, i.e., sentences that are more than an elementary constituent, yet do not contain essential elements like subject or predicate (e.g., 20150919.70: *Warum nicht schon früher?* (eng. Why not earlier?)). For the case that an interrupted sentence is later replaced (semantically) by a complete sentence, we used another new category called S+ which comprises the incomplete sentence S- and the replacing sentence S. An example is illustrated in Figure 3a. The speaker starts the sentence with *haben* (eng. to have) and then changes her mind to formulate the question differently. S+ nodes were either used for immediate corrections of incomplete sentences or in long sentences in which an interrupted sentence is taken up again at a later point. Figure 1c shows a case in which a sentence is incomplete because of a spontaneous insertion (node S/MO) and then it is taken up again, matching both the original sentence start and the insertion.

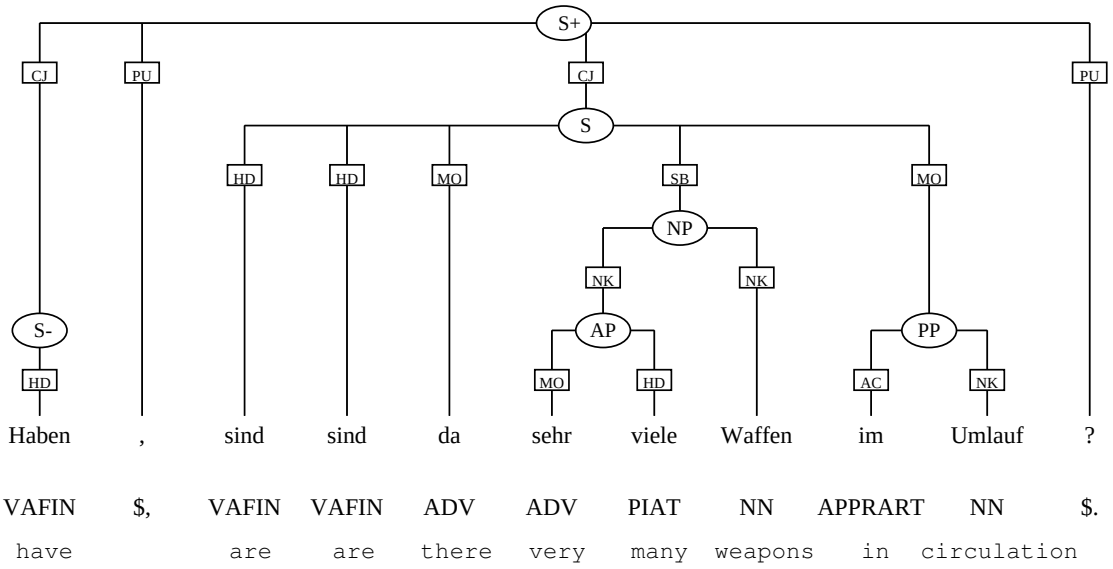
Preposition stranding. Another interesting phenomenon is the split of pronouns in German colloquial language which is also known as preposition stranding. For example, the German pronoun *dafür* (eng.: for that) is sometimes split into two words *da* and *für*. Preposition stranding is not covered by the TIGER syntax annotation guidelines, since it is mostly a phenomenon of spoken German. Figure 2 shows an example of such case. The tokens *Da* and *für* are used like the pronoun *dafür* to refer back to a previously mentioned noun phrase. To show the connection of the tokens *da* and *für*, it was decided to insert a prepositional phrase spanning both tokens. The stranded preposition (here: *für*) is attached with the label AC and the particle (here: *da*) with the label NK. Analogously to a pronoun, the prepositional phrase is attached to the S node as a modifier. Preposition stranding occurs rarely in the dataset, probably due to the fact that mostly experienced speakers took part in the interviews. In more colloquial or dialectal speech, this phenomenon will appear even more often.

Discourse markers. Slight modification was necessary regarding discourse markers. For TIGER, discourse markers are mainly response particles or interjections. We added the word *also* (eng: so, thus) to the set of discourse markers from the TIGER guidelines, when it is not used to introduce a conclusion.

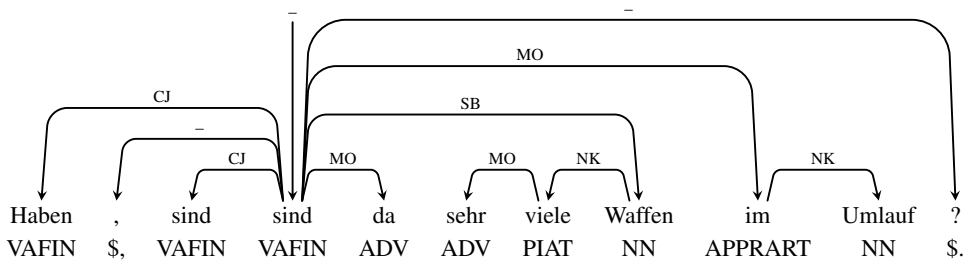
Punctuation. We introduce a new label category PU for punctuation and annotate it as part of the constituents of the sentence (see Figure 2 for an example). In the original TIGER, punctuation is not integrated with the rest of the sentence structure but instead attached to a virtual root node. Since punctuation often mark constituent boundaries and can provide clues for automatic systems we decided to integrate it into the syntactic structure. This also allows a cross-reference of their position with pauses or other speech phenomena in the audio track. In most cases, punctuation attachment is straight-forward and could be automated. However, in some rare cases it can be ambiguous, for example when multiple subordinate clauses are following one another or are embedded within each other. Since all punctuation is marked with the same label PU and part-of-speech tags, and no constituent is headed by punctuation, it can be removed automatically at any time without harming the syntactic structure of the sentence.

5. Conversion to Dependency Trees

To convert the constituency trees to dependency format we follow the conversion style presented in Seeker and Kuhn



(a) Constituency tree for the sentence 20150124.119.



(b) Dependency tree for the sentence 20150124.119.

Figure 3: Sentence 20150124.119: *Haben, sind sind da sehr viele Waffen im Umlauf?* (eng. Have, are are there a lot of weapons in circulation?).

(2012). The conversion tool `tiger2dep`[5] was developed to transform the full TIGER treebank to dependency structures and later extended to handle five out-of-domain test sets (Seeker and Kuhn, 2014). We follow this line of work and extend the conversion tool further to handle the interviews dataset.[6]

The conversion to dependency structures is performed bottom-up. Every constituency phrase gets assigned one head from the set of its children. The head selection is executed with a set of hand-written rules which take into consideration function labels, part-of-speech tags, and the order of the children. Since GRAIN-S closely follows TIGER guidelines for annotation we use the conversion rules designed for the full TIGER with small changes described below.

Punctuation. As described in Section 4.2., GRAIN-S deteriorates from the original TIGER in terms of punctuation. It introduces a new category PU and annotates punctuation as part of constituents (in TIGER all punctuation is part of a virtual root node). Since we want the final conversion to be as similar to the original TIGER as possible we remove all PU nodes from GRAIN-S before running `tiger2dep`. That way the default treatment of punctuation is applied, i.e., it is attached to the deepest common ancestor of the left and the right neighbor.

New categories. GRAIN-S introduces two new types of categories to deal with interrupted sentences: S- and S+. To deal with S- nodes we extend the set of head-finding rules and add S- with the same rules as S, i.e., it prefers heads with function label HD and does not impose any constrains on part-of-speech tag of the head.

S+ is treated as a coordinated sentence and does not need additional head-finding rules. Children of coordinated sentences in TIGER are marked with conjunct function CJ. `tiger2dep` converts such structures to dependencies by taking the first conjunct as the head of the coordination and creating a chain of the following conjuncts and coordinating conjunctions. We change this behavior slightly and add a constraint that the head conjunct can not be S- (unless there are only S- children). That way interrupted sentences become dependent on full sentences and not the other way around.

Figure 3 shows an example sentence from GRAIN-S containing a reparandum. The speaker starts by saying the verb *haben* (eng. to have) and then corrects the verb by saying *sind* (eng. to be). The constituency tree for the example

[5]Persistent identifier (PID): `http://hdl.handle.net/11022/1007-0000-0007-DFE2-F`

[6]We release `tiger2dep-1.3` together with the data.

	oov %	avg. oov / sent
TIGER test set	9.96	1.83
EuroParl	13.26	3.08
EuroParl-norm	4.49	1.04
novel	7.64	1.22
DVD manual	23.89	3.91
economy news	12.30	2.53
alpine stories	14.84	2.72
interviews all	5.97	1.07
interviews int.	7.74	1.10
interviews guest	5.46	1.06

(a) Ratios of out-of-vocabulary (oov) tokens and the average number of unknown tokens per sentence.

	1st & 2nd vs. 3rd	% of pers. pron.
TIGER train set	2052/9038	18.5/81.5
TIGER test set	341/1329	20.4/79.6
EuroParl	599/280	68.1/31.9
novel	181/326	35.7/64.3
DVD manual	227/45	83.5/16.5
economy news	9/29	23.7/76.3
alpine stories	564/242	70.0/30.0
interviews all	365/199	64.7/35.3
interviews int.	60/47	56.1/43.9
interviews guest	305/152	66.7/33.3

(b) Frequencies of personal pronouns by grammatical person and their fraction of all personal pronouns (PPER).

	# imperatives	% of verb forms
TIGER train set	114	0.1
TIGER test set	24	0.2
EuroParl	4	0.1
novel	5	0.4
DVD manual	183	15.2
economy news	0	0.0
alpine stories	0	0.0
interviews all	1	0.1
interviews int.	1	0.3
interviews guest	0	0

(c) Frequencies of imperatives (VVIMP,VAIMP) and their fraction of all verb forms (V*).

	# questions	% of all sentences
TIGER train set	657	1.6
TIGER test set	54	1.1
EuroParl	25	3.5
novel	93	17.6
DVD manual	0	0.0
economy news	0	0.0
alpine stories	26	2.5
interviews	77	12.3
interviews int.	68	38.4
interviews guest	9	2.0

(d) Frequencies of questions and their fraction of all sentences.

Table 2: Frequencies of specific linguistic phenomena selected by Seeker and Kuhn (2014) across all out-of-domain datasets. Statistics for interviews are presented for the whole dataset (all) and separately for utterances of guests and interviewers (int.).

sentence is presented in Figure 3a. The interrupted sentence is annotated with a node S-, the following repaired sentence with a node S, and the two nodes form an S+ constituent with function labels CJ. Figure 3b shows a result of conversion to a dependency structure. Token *sind* is the root of the sentence and the interrupted *haben* becomes its dependent.

Manual corrections. `tiger2dep` fails when it can not match any of the head-finding rules to a given constituent. This behavior is a design decision due to which all unexpected syntactic structures need manual inspection instead of being forced into a possibly flawed dependency structure.

Only 13 out of 626 GRAIN-S sentences failed to produce a dependency tree during the first run of the converter. The problems were mostly related to annotation inconsistencies or speech-specific phenomena. For example, in the sentence presented in Figure 3a the speaker repeated the verb *sind* twice. As a result node S has two children with the head function label HD and the converter needs additional information that the second one should be selected (see the result of conversion in Figure 3b).

6. Variation from other Domains

Out-of-domain test suites allow to investigate how well models generalize knowledge from training data and make use of it when applied to new genres. Since GRAIN-S keeps the same constituency and dependency representations as TIGER it can serve as an out-of-domain test set, expanding the existing TIGER-style test suite from Seeker and Kuhn (2014) (i.e., EuroParl, novels, DVD manuals, economy news, and alpine hiking stories) by interviews genre. To demonstrate in which aspects the new treebank is different from the ones there we compare frequencies of specific linguistic phenomena between interviews and other datasets. The specific phenomena were selected by Seeker and Kuhn (2014) and we refer the reader to their work for more details and analysis of differences across out-of-domain test sets.

Unknown word forms. Table 2a presents the frequency of out-of-vocabulary words when the training part of

TIGER serves as in-domain data. Interestingly, interviews have very small ratio of unknown word forms. Less than 6% of tokens do not occur in the training data, which is less than for any other genre. Our hypothesis is that since the interviews cover mostly political and social subjects they are topic-wise very close to TIGER, which consists of newspaper texts taken from the Frankfurter Rundschau. Moreover, the text from the interviews went through two manual creation stages, i.e., transcribing and textual unnormalization, which might have decreased the number of spelling errors and other written peculiarities. For example, the high number of out-of-vocabulary words for EuroParl comes from different spelling of umlauts and drops to 4.49% when they are normalized (see EuroParl-norm in the Table 2a).

1st & 2nd person vs. 3rd person. Since newspaper articles are written in a reporting style they contain less 1st and 2nd personal inflection than 3rd. Figure 2b gives a breakdown of personal pronouns in all the analyzed datasets.[7] Interviews differ a lot from TIGER – almost 65% of all personal pronouns is in 1st or 2nd person comparing to 18.5% and 20.4% for the training and testing parts of TIGER respectively. The most similar out-of-domain genre to the conversations is EuroParl, which is built from the proceedings of the European Parliament (68.1% of 1st and 2nd personal pronouns).

Imperatives and questions. Seeker and Kuhn (2014) compare newspaper texts with out-of-domain datasets by looking at the frequency of imperatives and questions. Figures 2c and 2d extend their statistics by the interviews dataset. We can notice that high frequency of imperatives is specific only for DVD manuals and it does not distinguish interviews from other genres. On contrast, questions are very common in interviews and question marks appear in 12.3% of all sentences, which puts them second after the novel dataset. Additionally, as expected from the nature of interviews questions are more frequent in utterances of journalists (38.4%) than of the guests (2%).

Baseline experiments. Test suites enable researchers to study different parsing strategies and adaptation methods in the out-of-domain setting. For future reference we provide baseline parsing results for the dependency-based part of GRAIN-S. Moreover, we compare performance of dependency parsers applied to interviews and other out-of-domain datasets to examine which of the domains poses biggest challenges to the parsing models.

Preprocessing. We use the same preprocessing pipeline as Seeker and Kuhn (2014), i.e., the CRF tagger MarMot (Mueller et al., 2013) for jointly predicting part-of-speech tags and morphological features and the lemmatizer from mate-tools[8] for lemmas. In all of the experiments the parsing models are trained on the TIGER train set annotated with preprocessing information via 5-fold jackknifing.

[7]The statistics differ from the ones reported by Seeker and Kuhn (2014). The authors by accident counted *Sie* form (a pronoun for politely addressing another person) as 3rd person. The biggest difference can be observed for DVD manuals which use a lot of *Sie* form to instruct the reader.

[8]https://code.google.com/archive/p/mate-tools/

	mate		IMSnPars	
	UAS	LAS	UAS	LAS
TIGER test set	90.35	88.17	92.16	90.41
EuroParl-norm	86.82	82.83	88.93	85.26
novel	88.42	83.98	90.83	86.81
DVD manual	83.20	79.31	85.65	82.15
economy news	83.67	79.98	84.19	81.54
alpine stories	84.78	81.39	89.21	86.52
interviews all	82.77	79.31	87.17	84.68
interviews int.	83.76	80.38	87.25	84.11
interviews guest	82.48	79.00	87.15	84.84

Table 3: Parsing performance for two dependency parsers: mate and IMSnPars. The models are trained on the training part of TIGER and applied to the out-of-domain test sets.

Parsers. Following Seeker and Kuhn (2014) we use the graph-based dependency parser from Bohnet (2010) which is a component of mate-tools. To compare this model with a more state-of-the-art tool, we take the BiLSTM-based graph-based parser from IMSnPars[9] described in Falenska and Kuhn (2019). The parser does not use lemmas and morphological tags. It builds token representations by concatenating pretrained word embeddings, character-based embeddings, part-of-speech tags, and ELMO deep contextualized word representations (Peters et al., 2018). For the pretrained word and ELMO representations we use the fastText vectors (Grave et al., 2018) and the German model provided by Che et al. (2018) respectively. We use default hyperparameters for both of the parsers and provide averages from three runs with different random seeds.

Results. Table 3 presents parsing performance in terms of unlabeled attachment score (UAS) and labeled attachment score (LAS) for both of the parsers. As expected parsing out-of-domain datasets is more difficult than the in-domain test set of TIGER. Similarly to the results of Seeker and Kuhn (2014), for models trained on newspaper articles the most challenging domains are DVD manuals and economy news.

Interestingly, for the mate parser interviews are as problematic as DVD manuals and the parser achieves only 79.31 LAS. Especially challenging are utterances of guests, for which the performance drops further to 79 LAS. One of the reasons might be the average length of sentences in this dataset. Guests use on average 19.5 tokens in one sentence which is more than in TIGER training part (17.78 tokens) and much more than in sentences spoken by interviewers (14.23 tokens).

IMSnPars clearly surpasses mate for both in-domain and out-of-domain setting. Its advantage ranges from 1.56 LAS for economy news up to 5.37 LAS for the interviews. Despite this advantage the interviews still pose a big challenge to the parser and are the third most difficult dataset to parse.

[9]https://github.com/AgnieszkaFalenska/IMSnPars

7. Conclusion and Discussion

We have presented GRAIN-S, i.e., an extension to the GRAIN release of the SFB732 Silver Standard Collection. The dataset comes with six interviews, each with (1) a merged version of TIGER-style constituency trees from two different annotators, (2) separate versions from the annotators, (3) a dependency conversion of the merged trees.[10] GRAIN-S follows the objectives of GRAIN by applying existing procedures and modifying them only where necessary to suit the out-of-domain setting. We have discussed the annotation process and the decisions needed to adapt the original guidelines to accommodate the primary data. Furthermore, we have presented a conversion to dependency syntax which is also based on the original guidelines and has been already applied to several domains of primary data.

Our dataset aims at bridging the gap between capabilities of standard text processing tools and the domain of spoken language. Dataset statistics showed that the interview genre differs in many aspects from other domains. Moreover, it poses a big challenge to state-of-the-art parsers because their performance drops significantly when applied to sentences from interviews.

The combination of different layers of annotation and metadata in GRAIN-S can serve as a valuable resource for linguistic research addressing questions combining speech- and text-processing, and even more distant topics such as gender bias. For example, Garimella et al. (2019) recently showed that statistical parsers perform differently on newspaper articles written by men and women. Since the latest release of TIGER contains information about the gender of the authors (Falenska et al., 2018), GRAIN-S can be used to test if similar patterns can be observed in spoken data.

8. Acknowledgements

This work was supported by the German Research Foundation (DFG) via SFB 732, projects D8 and INF.

9. Bibliographical References

Bohnet, B. (2010). Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China, August. Coling 2010 Organizing Committee.

Brants, S. and Hansen, S. (2002). Developments in the TIGER Annotation Scheme and their Realization in the Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 1643–1649, Las Palmas, Canary Islands - Spain, May. European Language Resources Association (ELRA).

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation*, 2(4):597–620.

[10]Persistent identifier (PID): `http://hdl.handle.net/11022/1007-0000-0007-DFDE-5`

Che, W., Liu, Y., Wang, Y., Zheng, B., and Liu, T. (2018). Towards Better UD Parsing: Deep Contextualized Word Embeddings, Ensemble, and Treebank Concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium, October. Association for Computational Linguistics.

Dannenberg, A., Werner, S., and Vainio, M. (2016). Prosodic and syntactic structures in spontaneous English speech. In *Speech Prosody 2016*, pages 59–63.

Dipper, S., Lüdeling, A., and Reznicek, M. (2013). NoSta-D: A Corpus of German Non-Standard Varieties. In Marcos Zampieri et al., editors, *Non-standard Data Sources in Corpus-based Research*, pages 69–76. Shaker.

Eckart, K. and Gärtner, M. (2016). Creating Silver Standard Annotations for a Corpus of Non-Standard Data. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, volume 16 of *BLA: Bochumer Linguistische Arbeitsberichte*, pages 90–96, Bochum, Germany.

Eckart, K., Riester, A., and Schweitzer, K. (2012). A Discourse Information Radio News Database for Linguistic Analysis. In Christian Chiarcos, et al., editors, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pages 65–75. Springer, Heidelberg.

Falenska, A. and Kuhn, J. (2019). The (Non-)Utility of Structural Features in BiLSTM-based Dependency Parsers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 117–128, Florence, Italy, July. Association for Computational Linguistics.

Falenska, A., Eckart, K., and Kuhn, J. (2018). Moving TIGER beyond Sentence-Level. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2203–2210, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Foth, K. A., Köhn, A., Beuck, N., and Menzel, W. (2014). Because Size Does Matter: The Hamburg Dependency Treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2326–2333, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Garimella, A., Banea, C., Hovy, D., and Mihalcea, R. (2019). Women's Syntactic Resilience and Men's Grammatical Luck: Gender-Bias in Part-of-Speech Tagging and Dependency Parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy, July. Association for Computational Linguistics.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3483–3487, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Hinrichs, E., Kübler, S., Naumann, K., Telljohann, H., and

Trushkina, J. (2004). Recent developments in linguistic annotations of the TüBa-D/Z treebank. In *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*, pages 51–62.

Mueller, T., Schmid, H., and Schütze, H. (2013). Efficient Higher-Order CRFs for Morphological Tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA, October. Association for Computational Linguistics.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Reyle, U. and Riester, A. (2016). Joint information structure and discourse structure analysis in an Underspecified DRT framework. In *Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue (JerSem)*, pages 15–24.

Riester, A. and Baumann, S. (2017). *The RefLex Scheme – Annotation Guidelines*, volume 14 of *SinSpeC. Working Papers of the SFB 732*. University of Stuttgart.

Riester, A., Brunetti, L., and De Kuthy, K. (2018). Annotation guidelines for Questions under Discussion and information structure. *Information Structure in Lesser-Described Languages: Studies in Prosody and Syntax*, pages 403–443.

Rohrer, C. and Forst, M. (2006). Improving coverage and parsing quality of a large-scale LFG for German. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 2206–2211, Genoa, Italy.

Schweitzer, K., Eckart, K., Gärtner, M., Falenska, A., Riester, A., Rösiger, I., Schweitzer, A., Stehwien, S., and Kuhn, J. (2018). German Radio Interviews: The GRAIN Release of the SFB732 Silver Standard Collection. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2887–2895, Miyazaki, Japan, May. European Languages Resources Association (ELRA).

Seeker, W. and Kuhn, J. (2012). Making Ellipses Explicit in Dependency Conversion for a German Treebank. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3132–3139, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Seeker, W. and Kuhn, J. (2014). An Out-of-Domain Test Suite for Dependency Parsing of German. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4066–4073, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Sekine, S. (1997). The Domain Dependence of Parsing. In *Fifth Conference on Applied Natural Language Processing*, pages 96–102, Washington, DC, USA. Association for Computational Linguistics.

Zeman, D., Nivre, J., et al. (2019). Universal Dependencies 2.5. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

10. Language Resource References

Fachbereich Informatik - Universität Hamburg. (2014). *The Hamburg Dependency Treebank*. 1.0.

Institut für deutsche Sprache und Linguistik - Humboldt-Universität zu Berlin, Sprachwissenschaftliches Institut - Ruhr-Universität Bochum. (2013). *NoSta-D Corpus of German Non-Standard Varieties*. 1.2.

Institut für Germanistik - Universität Potsdam, Institut für Maschinelle Sprachverarbeitung - Universität Stuttgart, Computerlinguistik - Universität des Saarlandes. (2007). *TIGER Corpus*. 2.1.

Institut für Maschinelle Sprachverarbeitung - Universität Stuttgart. (2012). *Diskurs-Informations-Radio-Nachrichten-Datenbank für Linguistische Analysen (DIRNDL)*. 1.0.

Institut für Maschinelle Sprachverarbeitung - Universität Stuttgart. (2018). *GRAIN Corpus – German Radio Interviews*. 1.0.

Seminar für Sprachwissenschaft - Universität Tübingen. (2018). *Tübingen Treebank of Written German / Newspaper Corpus (TüBa-D/Z)*. 11.