Treebank Profiling of Spoken and Written German

Erhard W. Hinrichs, Sandra Kübler
SfS-CL, University of Tübingen
Wilhelmstr. 19
72074 Tübingen, Germany
{eh,kuebler}@sfs.uni-tuebingen.de

Abstract

This paper profiles significant differences in syntactic distribution and differences in word class frequencies for two treebanks of spoken and written German: the TüBa-D/S, a treebank of transliterated spontaneous dialogs, and the TüBa-D/Z treebank of newspaper articles published in the German daily newspaper 'die tageszeitung' (taz). The approach can be used more generally as a means of distinguishing and classifying language corpora of different genres.

1 Introduction

It has often been pointed out that spoken language differs considerably from written texts. The discussion of such differences has typically focused on phenomena characteristic of spontaneous speech, such as false starts, hesitations, slips of the tongue, self-corrections, and elliptical utterances. Less attention has been paid to differences in syntactic distribution or differences in frequencies of word classes. The purpose of this paper is to conduct three case studies of the latter kind. The empirical basis for this investigation is provided by two treebanks of German - one of spoken and one of written language - that have been constructed at the University of Tübingen over the past ten years. The TüBa-D/S is a treebank of transliterated spontaneous dialogs that were collected as part of the Verbmobil project on speech-to-speech machine translation from German to English and to Japanese. The subject domain of these dialogs is primarily the scheduling of business meetings. The TüBa-D/Z is a treebank of a newspaper corpus. The corpus consists of issues of the German daily newspaper 'die tageszeitung' (taz) that appeared in April and May of 1999.

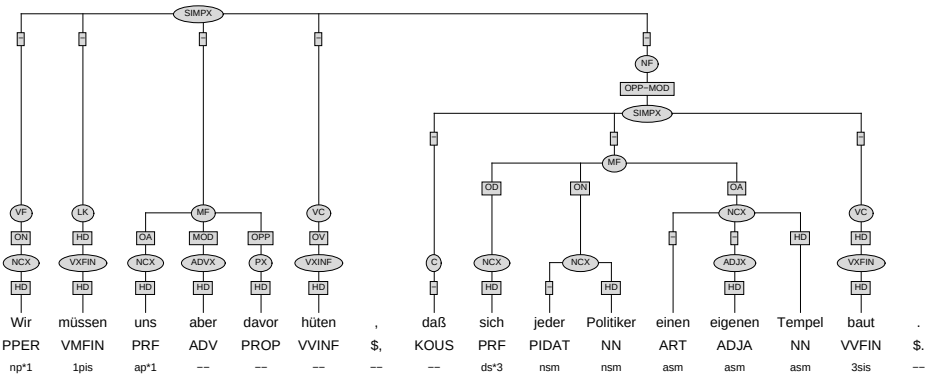

Figure 1: A TüBa-D/Z tree.

Both treebanks share virtually the same annotation scheme that has been documented by Stegmann et al. (2000) for TüBa-D/S and by Telljohann et al. (2003) for TüBa-D/Z. Part of speech assignment to lexical categories is provided by the Stuttgart-Tübingen tagset (STTS; Schiller et al. 1995), the standard inventory of parts-of-speech also used in the Negra (Skut et al., 1997) and Tiger treebank (Brants et al., 2002) developed independently of the Tübingen treebanks of German. Apart from phrasal and clausal annotations, the TüBa-D/S and the TüBa-D/Z treebanks include topological field annotations that identify the major grouping of constituents in the three different clause types of German.

The tree in Figure 1 illustrates the annotation scheme for sentence (1). The sentence (SIMPX) is grouped into the following topological fields (cf. section 3 for details): initial field (VF), left sentence bracket (LK), middle field (MF), verb complex (VC), and final field (NF). The finite verb constitutes the head (HD) of the clause. The grammatical relations annotated in the tree are: subject (ON), accusative object (OA), dative object (OD), verbal object (OV), prepositional object (OPP), modifier of the prepositional object (OPP-MOD), and modifier (MOD). The label OPP-MOD desribes a long-distance relationship, in which the subordinate clause modifies the prepositional object *davor*. The parts of speech and the morphological annotations are given below the lexical level.

(1) Wir müssen uns aber davor hüten, daß sich jeder Politiker
We need to us however from that prevent, that himself each politician
einen eigenen Tempel baut.
an own temple builds.
'But we have to prevent that every politician builds his own temple.'

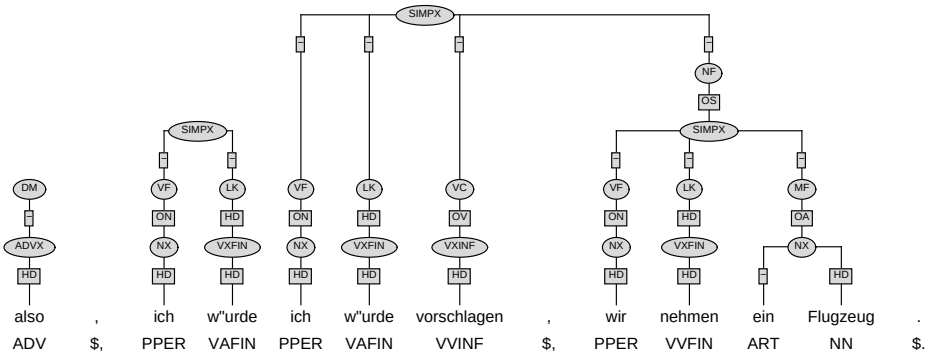

Figure 2: A sentence with a false start.

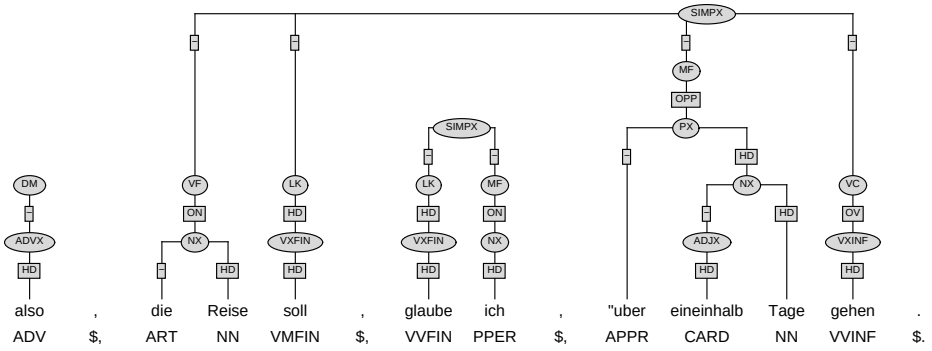

Figure 3: A parenthetical sentence.

In most cases, particularly for the treebank of written German, the annotation yields proper trees. However, there are exceptional cases where words or phrasal nodes remain unattached. Such cases include false starts (cf. Sentence (2) and Figure 2), parentheticals (cf. Sentence (3) and Figure 3), and elliptical utterances (cf. Sentence (4) and Figure 4).

(2) also ich würde ich würde vorschlagen, wir nehmen ein Flugzeug.
well I would I would suggest we take a plane.
'Well, I would suggest that we take a plane.'

(3) also, die Reise soll, glaube ich, über eineinhalb Tage gehen.
well the trip should, think I, over one and a half days go.
'Well, I think that the trip should be one and a half days long.'

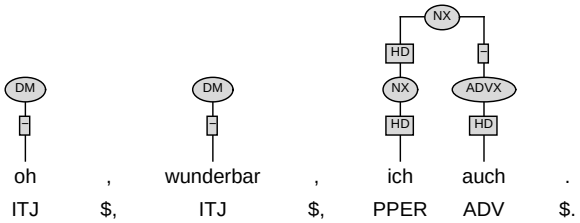

Figure 4: An elliptical utterance.

(4) oh, wunderbar, ich auch.
oh, wonderful, I also.
'Oh, wonderful, me, too.'

The treebanks were collected primarily as resources for research in computational linguistics. They have been used for the training of statistical parsers and for computational anaphora resolution. However, the treebanks are also a valuable resource for research in theoretical linguistics. In particular, they are of sufficient size to provide meaningful comparisons of spoken and written language. The TüBa-D/S consists of a total of 38,342 trees with a total number of 361,436 tokens. The TüBa-D/Z treebank currently consists of 22,087 trees with a total number of 381,558 tokens. The rich annotation scheme makes it possible to conduct fine-grained searches of the internal make-up of phrases and clauses as well as of their relative frequencies.

2 The Distribution of Noun Phrases

This section will compare the distribution of phrases and syntactic categories in the two treebanks and will focus on the distribution of noun phrases. Table 1 shows the distribution of noun phrases in the two treebanks.

The treebanks differ considerably in the relative frequency of different types of NPs. The term "definite NP" refers to NPs that start with a definite determiner, a demonstrative, or a possessive pronoun. In the newspaper treebank, such NPs are the most frequent among all NP types while in the treebank of spoken dialogs, they make up only 15.6% of all NPs. The distribution of pronouns (personal, possessive, and demonstrative) also differs significantly. In the TüBa-D/S (spoken) treebank, they make up almost half of all NPs while in the TüBa/D-Z (written) only 12.7% of all NP are pronouns. Although proper names are less frequent than definite NPs, indefinite NPs, and personal pronouns in both treebanks, their relative frequency is again significantly different, with proper names occurring three times more often in the TüBa/D-Z (written).

	TüBa-D/S (spoken):		TüBa-D/Z (written):	
number of NPs	86402		74935	
definite NPs	1348	15.6 %	28642	38.2 %
indefinite NPs	24832	28.7 %	23385	31.2 %
pronouns	41132	47.6 %	9506	12.7 %
proper names	2487	2.9 %	7153	9.6 %
relative pronouns	391	0.5 %	2746	3.7 %
reflexive pronouns	2792	3.2 %	2792	3.7 %
wh-questions	1284	1.5 %	711	1.0 %

Table 1: Distribution of NPs.

The term *indefinite* NP refers to all those NPs in the corpus that are not a member of any of the other classes listed in Table 1. While definite NPs outrank indefinite NPs in the newspaper corpus, the spoken language corpus exhibits a much different relative distribution, with indefinite NPs occurring almost twice as often as definite NPs.

The relative frequencies of NP types in the two corpora are indicative of the respective domains of the corpora. The topic structure in the dialogs is less cohesive than in newspaper texts since task-oriented dialogs such as appointment scheduling and travel planning involve discussion of different subtasks. The different distributions of definite and indefinite NPs reflect these differences. Indefinite NPs are typically used to introduce new discourse entities while definite NPs refer to entities that are "discourse-old". With relatively cohesive texts, it is to be expected that definite NPs become more frequent relative to indefinite NPs while the opposite is true for less cohesive dialogs.

The discourse function of pronouns is similar to that of definite NPs. In their anaphoric use, pronouns refer to events or entities previously introduced into the discourse. At first glance, the distribution of pronouns in the two treebanks (cf. Table 2) is rather surprising. However, a closer look at the types of pronouns used in the two corpora shows that first and second person pronouns as well as polite (morphologically third-person) pronouns are by far the most frequently used pronoun types in the dialog treebank. That the second person familiar pronouns (*du, ihr*) appear less frequently than the polite pronouns (*Sie, Ihnen*) is a direct reflection of the politeness requirements of the particular kind of dialogues. The primary use of pronouns in the dialog corpus is thus deictic rather than anaphoric. This is further highlighted by the fact that third person pronouns, which are typically used anaphorically (i.e. have a linguistic antecedent), make up only 10.5 % of all pronouns. By contrast, the deictic use of pronouns in the newspaper treebank is

	TüBa-D/S (spoken):		TüBa-D/Z (written):	
1st personal:	21880	53.2%	1957	20.6%
2nd person:	186	0.5%	83	0.9%
polite:	5933	14.4%	514	5.4%
3rd person (m/f):	314	0.8%	3194	33.6%
3rd person (n):	3999	9.7%	2139	22.5%
demonstratives	8935	21.7%	1518	16.0%

Table 2: Distribution of pronouns.

rather rare and is - we conjecture - largely restricted to direct speech environments such as quotations and headlines. Anaphoric third person pronouns make up the majority of all pronoun occurrences.

A related issue concerns the relative frequency of demonstrative pronouns in the treebanks. In the dialog treebank, demonstrative pronouns represent 21.7% of all pronouns while in the newspaper treebank only 16.0% are demonstratives.

3 Direct and indirect questions

The discussion in section 2 has focused on distributional properties that can be identified on the basis of POS information and syntactic annotation at the phrasal level. In this and the following section, we will utilize topological field information to consider more fine-grained distinctions in syntactic distribution between the two treebanks.

The theory of topological fields (Höhle, 1986) provides a layer of syntactic annotation between the level of individual phrases and the clause level. It is grounded in the placement of finite and non-finite verbs in different clause types of German. Consider the finite verb *wird* in (5) as an example.

(5) a. Peter wird das Buch gelesen haben.
Peter will the book read have.
'Peter will have read the book.'

b. Wird Peter das Buch gelesen haben?
Will Peter the book have read?
'Will Peter have read the book?'

c. dass Peter das Buch gelesen haben wird.
that Peter the book read have will.

		TüBa-D/S (spoken):		TüBa-D/Z (written):	
		counts	percentage	counts	percentage
C-Feld	nominal head	355	31.0%	458	69.3%
	any head	718	21.3%	803	68.0%
Vorfeld	nominal head	790	69.0%	203	30.7%
	any head	2648	78.7%	378	32.0%

Table 3: Distribution of nominal phrases in Vorfeld and C-Feld.

'... that Peter will have read the book.'

In non-embedded assertion clauses, the finite verb occupies the second position in the clause (V2), as in (5a). In yes/no questions, as in (5b), the finite verb appears clause-initially (V1) whereas in embedded clauses it appears clause finally (VL), as in (5c). Regardless of the particular clause type, any cluster of non-finite verbs, such as *gelesen haben* in (5a) and (5b) or *gelesen haben wird* in (5c), appears at the right periphery of the clause.

The positions of the verbal elements form the *Satzklammer* (sentence bracket) which divides the sentence into a *Vorfeld* (initial field), a *Mittelfeld* (middle field), and a *Nachfeld* (final field). The Vorfeld and the Mittelfeld are divided by the *linke Satzklammer* (left sentence bracket), which is realized by the finite verb or (in verb-final clauses) by a *C-Feld* (complementizer field). The *rechte Satzklammer* (right sentence bracket) is realized by the verb complex and consists of verbal particles or sequences of verbs. This right sentence bracket is positioned between the Mittelfeld and the Nachfeld.

Table 1 shows that wh-questions with nominal heads occur with roughly the same relative frequency in both treebanks. This seems rather surprising since one would expect that wh-questions would have a much higher occurrence in the TüBa-D/S treebank, considering the task-oriented dialogs it records. However, if one considers a more fine-grained classification of wh-questions into direct and embedded questions, then the distribution of these two question types is characteristically different. Topological field annotation enables us to distinguish between these two question types. Direct wh-questions are V2-clauses, in which the wh-phrase occurs in the Vorfeld while for indirect questions the wh-phrase appears in the C-Feld of a VL clause. As shown in Table 3, 69.0% of all wh-questions with a nominal head are direct questions in the dialog treebank while in the newspaper treebank only 30.7% are direct questions.

If one considers wh-questions with any head category, i.e. including also question words such as *wie, wo, wohin, woher, wann,* and *warum,* then the difference

	TüBa-D/S (spoken):	TüBa-D/Z (written):
wh-phrases in C-Feld	16.1%	10.1%
wh-phrases in Vorfeld	9.3%	1.7%

Table 4: Wh-phrases in C-Feld and Vorfeld.

in distribution between the two treebanks are even more apparent: in the dialog treebank, 78.7% of all wh-questions are direct questions while in the newspaper treebank, 32.0% are direct questions.

The distribution of nominal wh-questions and of all wh-questions among the two clause types is indicative of the two genres represented by the two treebanks, with direct questions naturally occurring more frequently in dialog data. It is also instructive to compare the percentages of wh-questions among all categories that occur in the C-Feld and the Vorfeld in the two treebanks.

In the dialog treebank, 16.1% of all subordinate clauses and 9.3% of all verb-second clauses are questions, as opposed to 10.1% for subordinate clauses and 1.7% for verb-second clauses in the newspaper corpus. Again, these relative frequencies of questions in the two treebanks is a reflection of the text types involved.

4 Syntactic Realization of the Vorfeld

Topological field annotation also provides the necessary information to study the distribution of sentence-initial constituents and their grammatical function in verb-second clauses in general. In the previous section we have already seen that the relative frequency of wh-questions in the Vorfeld differs considerably (9.3% in dialog corpus versus 1.7% in the newspaper corpus). Table 5 gives a summary of the relative frequencies for all grammatical functions in the Vorfeld for the two treebanks.

In both treebanks, approximately half of the Vorfeld constituents are subjects (nominal as well sentential subjects). Objects, on the other hand, occur rarely. We conjecture that the higher percentage of objects in the dialog corpus is due to the higher number of direct wh-questions that we discussed earlier.

Apart from subjects, modifiers make up the largest class of Vorfeld constituents. The labels MOD, V-MOD, and ON-MOD refer to the classes of sentential modifiers, verb phrase modifiers, and subject modifiers, respectively. The frequency rank of these modifiers differs in the two treebanks, with sentential modifiers outranking other modifiers by a large margin. Among sentential modifiers, 91.6% are realized as adverbial phrases in the dialog corpus, compared to 48.7%

	TüBa-D/S (spoken):		TüBa-D/Z (written):	
ON	14358	50.3%	11585	52.1%
MOD	7279	25.5%	3179	14.3%
V-MOD	2625	9.2%	3891	17.5%
OA	1682	5.9%	848	3.8%
PRED	1460	5.1%	495	2.2%
OS	191	0.7%	926	4.2%
ON-MOD	98	0.3%	279	1.3%
FRONTED FIELDS	23	0.01%	190	0.9%
OTHER	824	2.99%	749	3.7%

Table 5: Grammatical functions of Vorfeld constituents.

in the newspaper corpus. On the other hand, subordinate clauses make up 25.8% of all sentential modifiers in the newspaper corpus, but only 4.5 % in the dialog corpus. These differences in distribution are once again a reflection of the two genres involved: In the dialog corpus, discourse connectives such as *dann* ('then'), *deshalb* ('therefore') or *also* ('thus') figure prominently among adverbial phrases while the higher presence of clausal modifiers in the newspaper corpus is indicative of the higher frequency of hypotactic constructions in newspaper texts.

Another difference between the two corpora concerns the relative fequency of fronted topological fields. These are cases where non-finite verbs are fronted alone or together with complements or modifiers or where parts of the Mittelfeld appear in the Vorfeld. In the dialog corpus such highly complex constructions are extremely rare (0.01% all of Vorfeld realizations). While also rare in absolute terms (0.9%) in the newspaper corpus, they are much more frequent in the newspaper corpus than in the dialog corpus. The sentence in (6) shows a particularly complex example taken from the newspaper corpus where a verbal complex (*ausgenommen werden*) is fronted together with a Nachfeld PP-modifier (*von der neuen Steuer- und Sozialabgabenpflicht*). The annotation of this tree is shown in Figure 5. Examples such as (6) corroborate the claim of Müller (2003) that the Vorfeld need not be realized by a single constituent in German.

(6) Der Nordwest-Zeitung sagte Glogowski, ausgenommen werden von der
To the Nordwest-Zeitung said Glogowski, exempted be of the
neuen Steuer- und Sozialabgabenpflicht sollten Zeitungsträger,
new tax and social contributions should newspaper carriers,
Chorleiter oder Übungsleiter in Sportvereinen.
choirmasters or trainers in sports clubs.

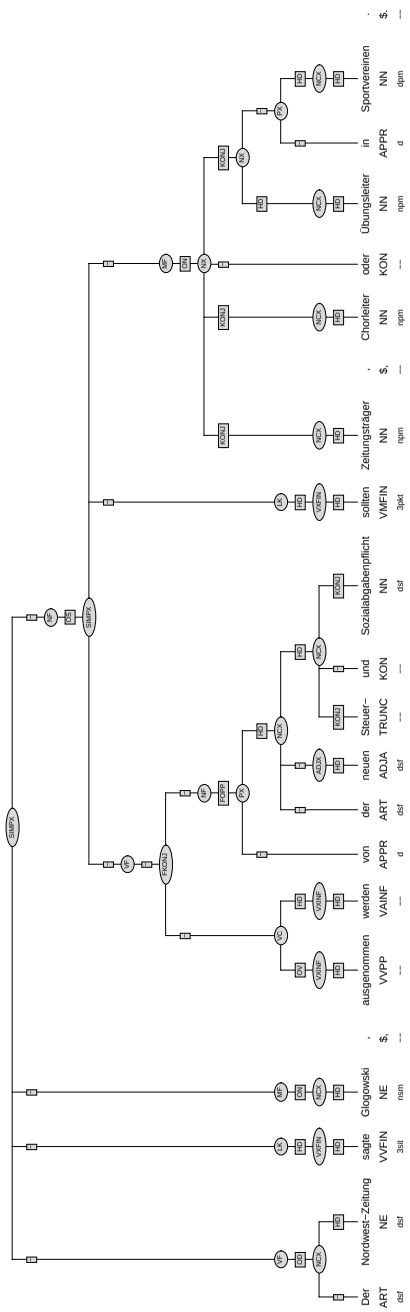

Figure 5: A TüBa-D/Z tree with a complex fronted field.

'Glogowski told the Nordwest-Zeitung that newspaper carriers, choirmasters, or trainers in sports clubs should be exempted from the new tax on wages and for social benefits.'

5 Conclusion and Outlook

We have presented a case study of profiling two treebanks from two rather different domains. While it is premature to draw more general conclusions from a single case study, we believe that the kinds of distributional tests presented here could be used more generally as a means of distinguishing and classifying language corpora of different genres. If successful, such profiling could be used to construct balanced corpora or identify subgenres within a heterogeneous corpus.

While the current study has relied on deep syntactic annotation of a corpus in the form of a treebank, it is important to note that the type of distributional information that we have profiled for the two treebanks can also be obtained by more shallow methods of analysis. Müller (2005) has shown that topological field information can be effectively combined with identification of so-called chunks, i.e. non-recursive syntactic phrases. Müller and Ule (2002) have developed a finite-state parser for German that has been used to automatically parse and partially annotate a very large corpus of German[1].

In sum, thanks to recent advances in computational linguistics, it is now possible to study interesting grammatical phenomena on the basis of large-scale, linguistically annotated corpora and to profile the distribution of grammatical functions and categories.

References

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER treebank. In Hinrichs, E. and Simov, K., editors, *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002)*, pages 24–41, Sozopol, Bulgaria.

Höhle, T. (1986). Der Begriff "Mittelfeld", Anmerkungen über die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses 1985*, pages 329–340, Göttingen, Germany.

[1] Cf. `http://www.sfs.uni-tuebingen.de/en_tuepp.shtml` for details on the corpus.

Müller, F. H. (2005). *A Finite-State Approach to Shallow Parsing and Grammatical Functions Annotation of German*. PhD thesis, Seminar für Sprachwissenschaft: University of Tübingen.

Müller, F. H. and Ule, T. (2002). Annotating Topological Fields and Chunks–and Revising POS Tags at the Same Time. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 695–701.

Müller, S. (2003). Mehrfache Vorfeldbesetzung. *Deutsche Sprache*, 30(1):29–.62.

Schiller, A., Teufel, S., and Thielen, C. (1995). Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart and Universität Tübingen.

Skut, W., Krenn, B., Brants, T., and Uszkoreit, H. (1997). An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP)*, Washington, D.C.

Stegmann, R., Telljohann, H., and Hinrichs, E. W. (2000). Stylebook for the German Treebank in Verbmobil, Verbmobil Report 239.

Telljohann, H., Hinrichs, E. W., and Kübler, S. (2003). *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Germany.