A Testsuite for Testing Parser Performance on Complex German Grammatical Constructions

Sandra Kübler
Indiana University
Department of Linguistics
skuebler@indiana.edu

Ines Rehbein
Universität des Saarlandes
Computational Linguistics
rehbein@coli.uni-sb.de

Josef van Genabith
Dublin City University
School for Computing
josef@computing.dcu.ie

1 Introduction

Traditionally, parsers are evaluated against gold standard test data. This can cause problems if there is a mismatch between the data structures and representations used by the parser and the gold standard. A particular case in point is German, for which two treebanks (TiGer and TüBa-D/Z) are available with highly different annotation schemes for the acquisition of (e.g.) PCFG parsers. The differences between the TiGer and TüBa-D/Z annotation schemes make fair and unbiased parser evaluation difficult [7, 9, 12]. The resource (TEPACOC) presented in this paper takes a different approach to parser evaluation: instead of providing evaluation data in a single annotation scheme, TEPACOC uses comparable sentences and their annotations for 5 selected key grammatical phenomena (with 20 sentences each per phenomena) from both TiGer and TüBa-D/Z resources. This provides a 2 times 100 sentence comparable testsuite which allows us to evaluate TiGer-trained parsers against the TiGer part of TEPACOC, and TüBa-D/Z-trained parsers against the TüBa-D/Z part of TEPACOC for key phenomena, instead of comparing them against a single (and potentially biased) gold standard. To overcome the problem of inconsistency in human evaluation and to bridge the gap between the two different annotation schemes, we provide an extensive error classification, which enables us to compare parser output across the two different treebanks.

In the remaining part of the paper we present the testsuite and describe the grammatical phenomena covered in the data. We discuss the different annotation strategies used in the two treebanks to encode these phenomena and present our error classification of potential parser errors.

2 TEPACOC - The Testsuite

TEPACOC contains 200 sentences carefully selected from two German treebanks. The sentences cover five complex grammatical constructions which are extremely difficult for a PCFG parser to process:

1. PP Attachment: Noun (PPN) vs. Verb Attachment (PPV)
2. Extraposed Relative Clauses (ERC)
3. Forward Conjunction Reduction (FCR)
4. Subject Gap with Finite/Fronted Verbs (SGF)
5. Coordination of Unlike Constituents (CUC).

PP attachment is the canonical case of structural ambiguity and constitutes one of the major problems in (unlexicalised) parsing, since disambiguation often requires lexical rather than structural information [5]. The testsuite allows us to investigate which of the different encoding strategies in the two treebanks is more successful in resolving PP attachment ambiguities.

The second construction we included in TEPACOC are extraposed relative clauses. According to Gamon et al. [2], who present a case study in German sentence realisation, 35% of all relative clauses in a corpus of German technical manuals are extraposed, while in a comparable corpus of English technical manuals less than one percent of the relative clauses have been subject to extraposition. This shows that extraposed relative clauses are a frequent phenomenon in German and worthwhile to be considered for parser evaluation.

Coordination is a phenomenon which poses a great challenge not only to statistical parsing but also to linguistic theories in general (see for example [6, 13, 11, 15] for a discussion on different types of coordination in LFG, HPSG, GPSG and CCG). Harbusch and Kempen [4] present a corpus study on the TiGer treebank (Release 2), where they investigate cases of clausal coordination with elision. They found 7196 sentences including clausal coordinations, out of which 4046 were subject to elisions. 2545 out of these 4046 sentences proved to be Forward Conjunction Reduction, and 384 sentences contained Subject Gaps with Finite/Fronted Verbs. We included FRC and SGF as the most frequent forms of non-constituent coordination in the testsuite. The TiGer treebank (Release 2) contains 381 sentences with at least one CUC, which means that coordination of unlike constituents are as frequent as SGF. Additionally, we choose CUC to be part of the TEPACOC because, from a linguistic point of view, they are quite interesting and put most linguistic theories to the test. The testsuite is available as a list of sentence numbers referring to the original treebanks so that interested parties can extract the sentences.[1]

[1] `http://jones.ling.indiana.edu/~skuebler/tepacoc`

2.1 Data Sources: TiGer and TüBa-D/Z

The data for the corpus comes from two different sources: the TIGER treebank (Release 2) [1] and the TüBa-D/Z (Release 3) [16]. Both treebanks contain German newspaper text and are annotated with phrase structure and dependency (functional) information. While both treebanks employ the same POS Tag Set (STTS) [14], the number of category labels and grammatical function labels varies. The most important differences between the two treebanks are: (1) the annotation in TIGER is rather flat compared to the more hierarchical annotation in TüBa-D/Z, (2) TIGER does not annotate unary branching, (3) TüBa-D/Z annotates topological fields, and (4) long distance dependencies in TiGer are expressed via crossing branches while in TüBa-D/Z, the same phenomenon is expressed with the help of grammatical function labels.

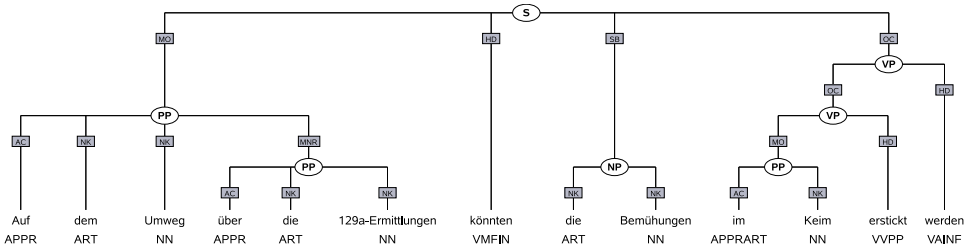
For each of the 5 grammatical phenomena listed above, we selected 20 sentences with a lenght ≤ 40 from each TIGER and TüBa-D/Z. This results in a test set of 200 sentences, 100 from each treebank. Below we give a survey of the testsuite: we describe the annotation of the phenomena in TEPACOC and discuss the different annotation decisions made in TIGER and TüBa-D/Z. The differences in treebank design do not support a systematic description of different error types like e.g. span errors, attachment errors or grammatical function label errors, as the same phenomenon might be encoded with the help of GF labels in one treebank and by using attachment in the other treebank. Therefore, we present a descriptive error classification scheme based on empirical data, capturing all potential parser errors on the specific grammatical phenomena.

2.2 PP Attachment: Noun (PPN) vs. Verb Attachment (PPV)

PP attachment is one of the problems discussed most in parsing since a correct attachment often requires lexical rather than purely structural information. In TIGER, noun attachment results in a flat tree structure in which the PP is attached on the same level as the head noun, while verb attachment has the PP grouped under the VP or the S node. Both NP and PP attachment are present in the TiGer example (1).[2]

In TüBa-D/Z, NP postmodifiers are attached on a higher level once the NP is grouped. For verb attachment the PP is directly attached to the governing topological field, the functional label shows whether it is considered a prepositional object (OPP), an optional prepositional object (FOPP), an unambiguous verbal modifier (V-MOD), or an ambiguous one (MOD). (2) shows a TüBa-D/Z representation of NP and VP attachment.

[2] Some of the examples have been shortened for readability.



(1) Auf dem Umweg **über die 129a-Ermittlungen** könnten die Bemühungen **im Keim** erstickt
By the detour via the 129a-investigations could the efforts in the bud nipped
werden.
be.
"With the 129a investigations, the efforts of the autonomouns activists could be nipped in the bud."

Error description		TIGER / TüBa-D/Z
(A)	correct GF & correct head of PP, span incorrect	
(B)	correct span, incorrect GF	
(C)	incorrect span, incorrect GF	
(D)	wrong attachment	

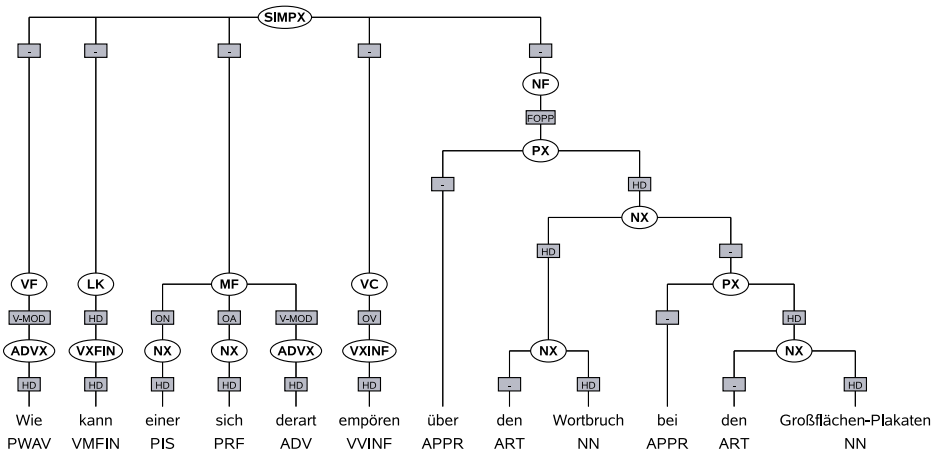
Table 1: Error classification for PP attachment

2.2.1 Error Classification (PPN vs. PPV)

We consider PP attachment parsed correctly if the PP is recognized correctly and if it is attached correctly with the correct grammatical function (Table 1). In TüBa-D/Z, extraposed PPs that are extracted from a preceding NP are not attached directly to the NP, their attachment is indicated in the grammatical function label. If an extraposed PP is attached incorrectly, the GF label is incorrect. In such cases, error code D must be used.

2.3 Extraposed Relative Clauses (ERC)

Extraposed relative clauses in German are treated as adjuncts to the head noun they modify, but there is no agreement in the literature whether they are base-generated locally [3] or get their final position through movement [10]. In TIGER, relative clauses are attached to the mother node of the head noun, which results in crossing branches for extraposed clauses, as in (3). The relative clause has the categorial node label S and carries the GF label RC. The relative pronoun is attached directly to the S node.



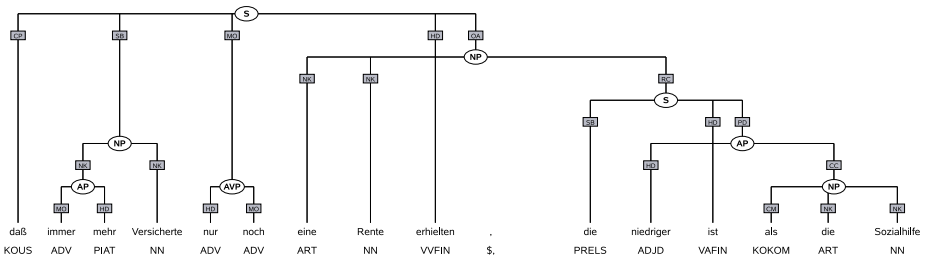
(2) Wie kann einer sich derart empören **über den Wortbruch bei den**
How can one *refl.* so revolt about the breach of promise concerning the
Großflächen-Plakaten?
large-scale posters?
"How can someone bristle at the breach of promise concerning the large-scale posters?"

In TüBa-D/Z, the extraposed relative clause is located in the final field (NF) and is associated with the node label R-SIMPX. The grammatical function label references the head noun modified by the relative clause, as in (4). The relative pronoun is embedded inside of an NP (NX) which is attached to a C node (complementizer for verb-final sentences).

2.3.1 Error Classification (ERC)

We consider an extraposed relative clause parsed correctly if the clause has been identified by the parser as a relative clause and is associated with the correct head noun, and if the phrase boundaries have been recognized correctly. Due to differences in annotation, here we have to adapt the error analysis to the annotation scheme of each treebank. Table 2 shows our error classification for extraposed relative clauses with an error specification for each treebank.

In TIGER, the grammatical function label carries the information that the clause is a relative clause while in TüBa-D/Z, the same information is encoded in the categorial node label. Therefore, error description (A) corresponds to a function label error in TIGER and to a categorial node label error in TüBa-D/Z. The relationship between the relative clause and its head noun is expressed through attachment in TIGER and by the use of a GF label in TüBa-D/Z. Therefore (B) is caused by a



(3) ... dass immer mehr Versicherte nur noch eine **Rente** erhielten, **die niedriger ist als die Sozialhilfe**
... that always more insurants just still a pension would receive, which lower is than the social welfare

"... that more and more insurants receive a pension lower than social welfare"

Error description	TIGER	TüBa-D/Z
(A) Clause not recognized as rel. cl.	Grammatical function incorrect	SIMPX label instead of R-SIMPX
(B) Head noun incorrect	Attachment error	Grammatical function incorrect
(C) Clause not recognized	Clause not recognized	Clause not recognized
(D) Clause boundaries not correct	Span error	Span error

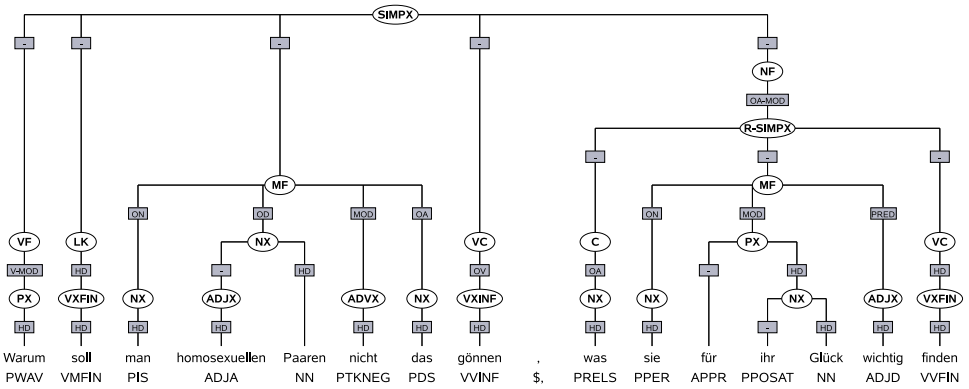
Table 2: Error classification for extraposed relative clauses

wrong attachment decision in TIGER and by a GF label error in TüBa-D/Z. For (C) the parser fails to identify the relative clause alltogether. This is usually caused by a POS tagging error, i.e. when the parser fails to assign the correct POS tag to the relative pronoun. (D) applies to both annotation schemes: here, the main components of the clause have been identified correctly but the phrase boundaries are slightly wrong.

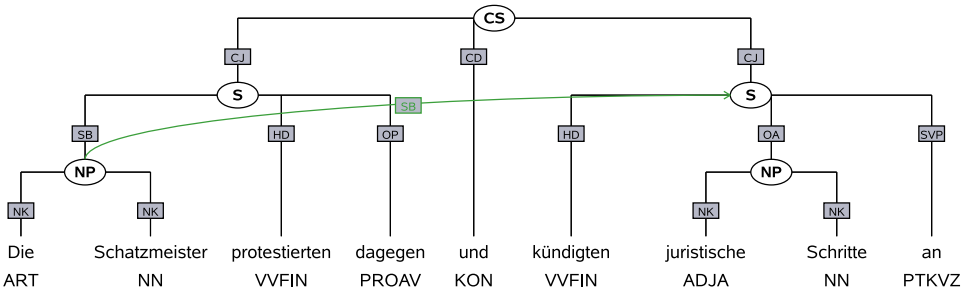
2.4 Forward Conjunction Reduction (FCR)

Forward Conjunction Reduction is a form of coordination in which both conjuncts contain a main verb and its arguments, and in which the conjuncts share the left peripheral context. In TIGER, the coordination is on sentence level, as in (5). The left peripheral context and the first conjoined verb phrase are grouped as a clause (S), and the second conjunct is projected to an elliptical clause. Both clauses are then coordinated. The role of the left peripheral context in the second clause is annotated via a secondary edge.

In TüBa-D/Z, the coordination is on the level of field combinations, as in (6).



(4) Warum soll man homosexuellen Paaren nicht **das** gönnen, **was sie für ihr Glück wichtig finden?**
Why shall one homosexual couples not that grant, which they for their luck important find?
"Why shouldn't homosexual couples be granted what they think is important to their happiness."

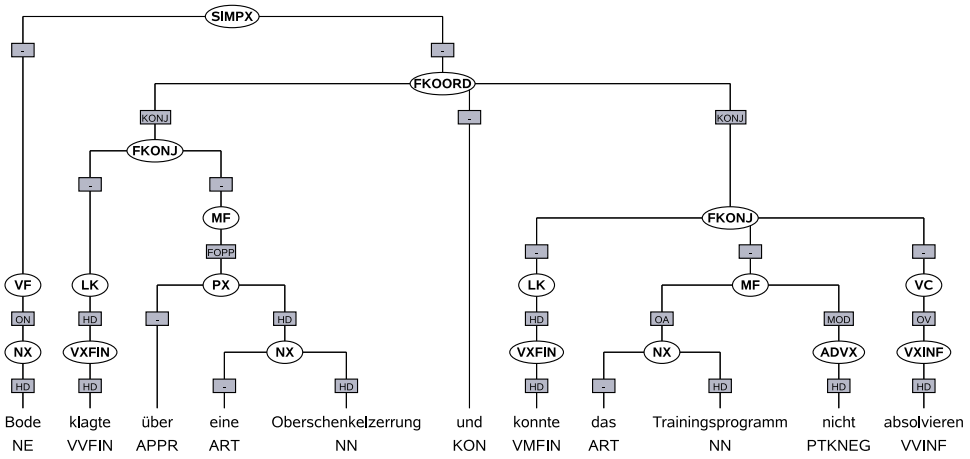


(5) Die Schatzmeister protestierten dagegen und kündigten juristische Schritte an.
The treasurers protested against it and announced legal action *verb part.*
"The treasurers protested and announced, they would take legal action."

As a consequence of the field model, the left peripheral context constitutes the initial field (VF) and is attached only once the coordination is grouped. Within the coordination, each conjunct is a combination of the verbal field (LK or VC) and its arguments (MF).

2.4.1 Error Classification (FCR)

We consider the FCR parsed correctly if the parser has identified the coordination, has assigned the subject label to the appropriate node, and if no other node in the



(6) Bode klagte über eine Oberschenkelzerrung und konnte das Trainingsprogramm nicht
Bode complained about a strain of the thigh and could the training regime not
absolvieren.
finish.
"Bode complained about a strain of the femural muscle and could not finish the training."

first or second constituent has been associated with the subject label. Here the annotation schemes allow us to use the same error specification for both treebanks (Table 3).

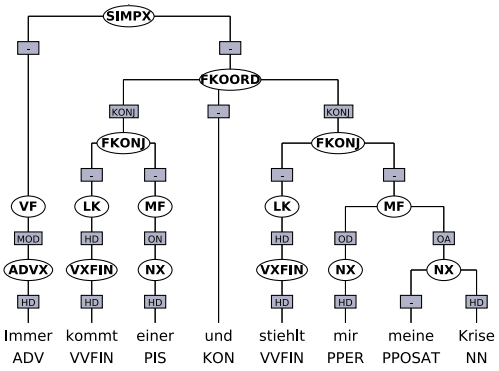
Error description		TIGER / TüBa-D/Z
(A)	Parser incorrectly annotates subject in one of the constituents	
(B)	Parser fails to identify subject	
(C)	Coordination not recognized	
(D)	Second subject in first conjunct	
(E)	Span error	(only in TüBa-D/Z)

Table 3: Error classification for forward conjunction reduction

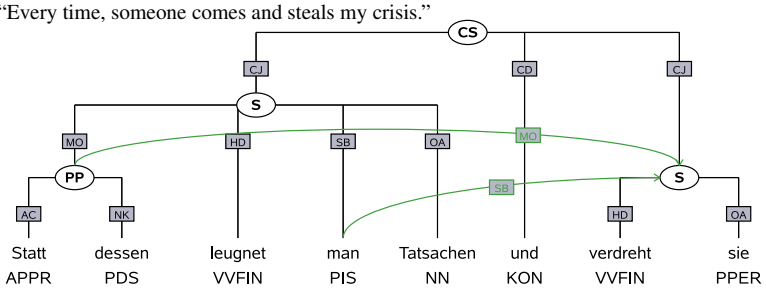
2.5 Subject Gap with Fronted/Finite Verbs (SGF)

Next, we discuss the case of asymmetric coordination where the subject of the left conjunct is realized in the middle field, while in the right conjunct the subject is missing. In TüBa-D/Z, subject gapping is treated as a complex coordination of fields (FKOORD), as in (7). The subject is realized in the middle field of the first constituent and has the functional label ON (nominative object). Both constituents are associated with the functional label FKONJ (conjunct with more than one field).

In TIGER, subject gaps with fronted/finite verbs are encoded as a coordination



(7) Immer kommt **einer** und stiehlt mir meine Krise.
Always comes someone and steals me my crisis.
"Every time, someone comes and steals my crisis."



(8) Statt dessen leugnet **man** Tatsachen und verdreht sie.
Instead denies one facts and twists them.
"Instead, the facts are denied and twisted."

of sentences (CS), as in (8). As in TüBa-D/Z, the subject is realized in the first constituent and can be identified by the grammatical function label SB (subject). With the help of labeled secondary edges (SB), TIGER encodes explicitly that the subject of the first constituent should also be interpreted as the subject of the second constituent.

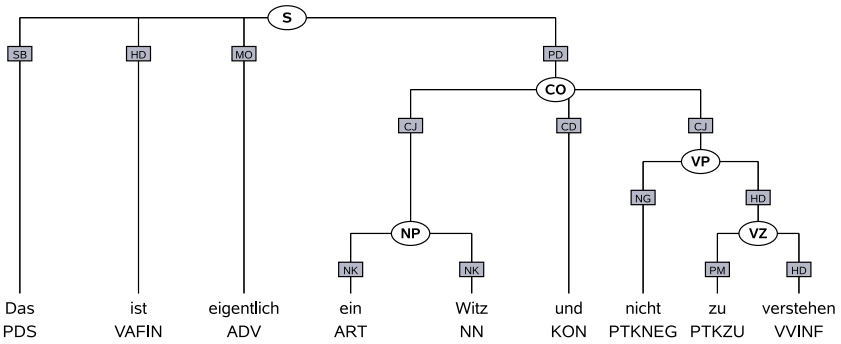
2.5.1 Error Classification (SGF)

We consider the subject gap construction parsed correctly if the parser has identified the coordination, has assigned the subject label to the right node in the first constituent, and no other node in the first or second constituent has been associated with the subject label. Here, the annotation schemes allow us to use the same error specification for both treebanks (Table 4).

Error description		TIGER / TüBa-D/Z
(A)	Parser incorrectly annotates subject in second conjunct	
(B)	Parser fails to identify subject in first conjunct	
(C)	Coordination not recognized	
(D)	Parser annotates additional subject in first conjunct	
(E)	Parser fails to identify the verb in the sentence	

Table 4: Error classification for subject gap with fronted/final verb

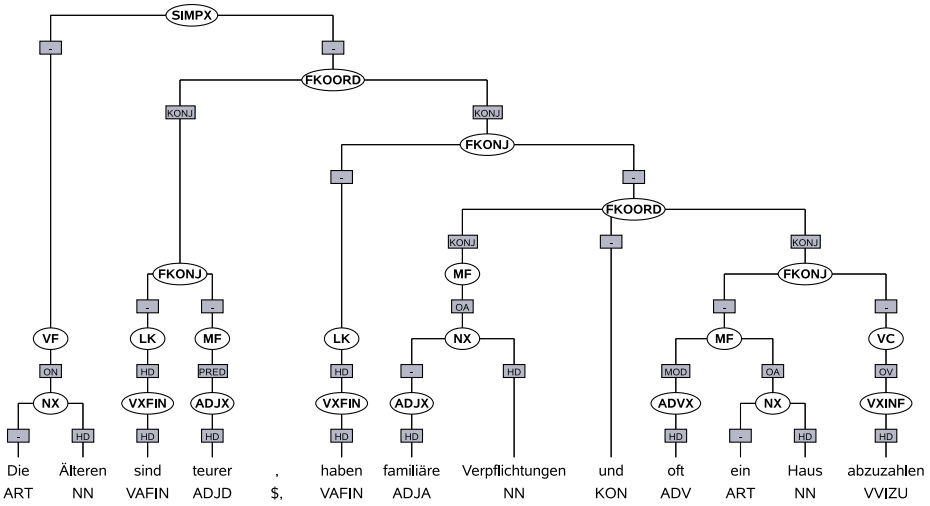
2.6 Coordination of Unlike Constituents (CUC)



(9) Das ist eigentlich ein Witz und nicht zu verstehen.
This is actually a joke and not to understand.
"This actually is a joke and hard to understand."

This section covers three types of coordinations of unlike constituents: VPs coordinated with adjectival phrases (AP), VPs coordinated with NPs, and clauses (S) coordinated with NPs. Here, we will concentrate on the second type since it shows the greatest differences between the two annotations schemes: in TIGER, the coordination is rather straightforward, the VP and the NP project to a coordinated phrase (CO), as in (9). Since the functional labels for the conjuncts (CJ) describe their conjunct status and the functional label of the coordination is the same as that associated with verb phrase (OC), the annotation does not contain explicit information which grammatical function the NP performs in the clause.

In TüBa-D/Z, the coordination is on the field level and the VP is represented as a combination of the verbal field and the middle field (MF), as in (10). The NP is projected to the MF, too, before both conjuncts are coordinated. In this case, the individual grammatical functions are retained in the constituents under the MFs.



(10) Die Älteren sind teurer, haben familiäre Verpflichtungen und oft ein Haus abzuzahlen.
The elderly are more expensive, have familial committments and often a house to repay.
"The elderly are more expensive, have family committments and often have to pay off a house."

2.6.1 Error Classification (CUC)

Since the two annotation schemes differ drastically in the annotation of coordinations of unlike constituents, we decided to use a correct/incorrect distinction only. A CUC is considered correct if the constituents are recognized with correct spans and correct heads.

2.7 Discussion

The phenomena described above are hard nuts to crack for a PCFG parser, and occur quite frequently in German text. The testsuite allows us (1) to assess the performance of constituent-based parsers on the grammatical constructions described above, and (2) to accomplish a fine-grained investigation on the impact of specific treebank design decisions on parser performance for specific syntactic phenomena. We showed that the two German treebanks choose different ways to encode the same syntactic phenomena. TEPACOC provides an additional means to answer questions like the following:

- What is more suitable to disambiguate PP attachment: the flat trees in TiGer or the more hierarchical annotation in TüBa-D/Z?
- Do topological fields support the identification of coordinations?

- Which of the two annotation strategies is more adequate to resolve non-local dependencies, as in ERC, FCR and SGF constructions?

Kübler et al. [8] put the TEPACOC to the test and compare results for constituent-based and dependency-based automatic evaluation measures with a manual evaluation on the TEPACOC sentences. They show that constituent-based evaluation measures are highly biased towards the more hierarachical annotation scheme of the TüBa-D/Z, while a dependency-based evaluation gives better results for labelled accuracy for parsers trained on the flat structures of the TiGer treebank. The dependency-based evaluation is backed up by a manual evaluation on the TEPACOC sentences, which sheds some light on the underlying reasons for the difference in parser performance on the two treebanks: (1) TiGer benefits from the flat annotation which makes it more transparent for the parser to detect constructions like ERC, FCR and SGF; (2) TüBa-D/Z suffers from the more hierarchical structure where relevant clues are embedded too deep in the tree for the parser to make use of it; (3) the additional layer of topological fields in TüBa-D/Z increases the number of possible attachment positions (and so of possible errors); and (4) topological fields reduce the number of rules in the grammar and improve the learnability especially for small training sets.

3 Conclusion and Future Work

We presented TEPACOC, a corpus for testing parser performance on complex grammatical constructions. TEPACOC covers five grammatical phenomena and provides a well-defined error categorisation which enables us to observe the influence of treebank design on specific grammatical constructions and so gain valuable insights for the future development and standardisation of language resources.

At the moment, TEPACOC includes German data only, but it can easily be extended to other languages. It is understood that the limited size of the testsuite does challenge the representativeness of the results. Therefore, TEPACOC should be used in addition to other evaluation metrics, providing an additional means to assess parser performance on a linguistic level and enabling us to compare results across different annotation schemes and languages.

References

[1] Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. The TIGER Treebank. In *Proceedings of TLT 2002*, Sozopol, Bulgaria, 2002.

[2] Michael Gamon, Eric Ringger, Zhu Zhang, Robert Moore, and Simon Corston-Oliver. Extraposition: a case study in German sentence realization. In *Proceedings of COLING 2002*, Morristown, NJ, USA, 2002.

[3] Hubert Haider. Downright down to the right. *Linguistik Aktuell*, 11:245–271, 1996.

[4] Karin Harbusch and Gerard Kempen. Clausal coordinate ellipsis in German: The TIGER Treebank as a source of evidence. In *Proceedings of NODALIDA 2007*, Tartu, Estonia, 2007.

[5] D. Hindle and M. Rooth. Structural ambiguity and lexical relations. *Computational Linguistics*, 19:103–120, 1993.

[6] Ron M. Kaplan and John Maxwell. Constituent coordination in lexical-functional grammar. In *Proceedings of COLING 1989*, Budapest, Hungary, 1989.

[7] Sandra Kübler. How do treebank annotation schemes influence parsing results? Or how not to compare apples and oranges. In *Proceedings of RANLP 2005*, Borovets, Bulgaria, 2005.

[8] Sandra Kübler, Wolfgang Maier, Ines Rehbein, and Yannick Versley. How to compare treebanks. In *Proceedings of LREC 2008*, Marrakech, Morocco, 2008.

[9] Wolfgang Maier. Annotation schemes and their influence on parsing results. In *Proceedings of the COLING/ACL 2006 Student Research Workshop*, Sydney, Australia, 2006.

[10] Gereon Müller. On extraposition and successive cyclicity. In *Syntax: Critical Concepts in Linguistics*, volume III. Routledge, 2006.

[11] Carl Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar*. University of Chicago Press and CSLI Publications, Chicago, Illinois, 1994.

[12] Ines Rehbein and Josef van Genabith. Treebank annotation schemes and parser evaluation for German. In *Proceedings of EMNLP/CoNLL 2007*, Prague, Czech Republic, 2007.

[13] Ivan A. Sag, Gerald Gazdar, Thomas Wasow, and Steven Weisler. Coordination and how to distinguish categories. Technical report, CSLI-84-3. Center for the Study of Language and Information, Sumford, California, 1984.

[14] Anne Schiller, Simone Teufel, and Christine Thielen. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart and Universität Tübingen, 1995.

[15] Mark Steedman. Dependency and coordination in the grammar of Dutch and English. *Language*, (61):523–568, 1985.

[16] Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, and Heike Zinsmeister. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Universität Tübingen, Germany, 2005.