Creating and Exploiting a Resource of Parallel Parses

Christian Chiarcos[*] and **Kerstin Eckart**[**] and **Julia Ritz**[*]

[*] Collaborative Research Centre 632
"Information Structure"
Universität Potsdam
{chiarcos|jritz}@uni-potsdam.de

[**] Collaborative Research Centre 732
"Incremental Specification in Context"
Universität Stuttgart
eckartkn@ims.uni-stuttgart.de

Abstract

This paper describes the creation of a resource of German sentences with multiple automatically created alternative syntactic analyses (parses) for the same text, and how qualitative and quantitative investigations of this resource can be performed using ANNIS, a tool for corpus querying and visualization. Using the example of PP attachment, we show how parsing can benefit from the use of such a resource.

1 Introduction

In this paper, we describe the workflow and the infrastructure to create and explore a corpus that contains multiple parses of German sentences. A corpus of alternative parses created by different tools allows us to study structural differences between the parses in a systematic way.

The resource described in this paper is a collection of German sentences with *-ung* nominalizations extracted from the SDeWaC corpus (Faaß et al., 2010), based on the DeWaC web corpus (Baroni and Kilgarriff, 2006). These sentences are employed for the study of lexical ambiguities in German *-ung* nominalizations (Eberle et al., 2009); e.g., German *Absperrung*, derived from *absperren* 'to block', can denote an event ('blocking'), a state ('blockade') or an object ('barrier'). Sortal disambiguation, however, is highly context-dependent, and reliable and detailed analyses of the linguistic context are crucial for a sortal disambiguation of these nominalizations.

More reliable and detailed linguistic analyses can be achieved, for example, by combining the information produced by different parsers: On the basis of qualitative and quantitative analyses, generalized rules for the improvement of the respective parsers can be developed, as well as rules for the mapping of their output to a tool-independent representation, and weights for the parallel application and combination of multiple parsers. This approach has been previously applied to morphological and morphosyntactic annotations (Borin, 2000; Zavrel and Daelemans, 2000; Tufiş, 2000), but only recently to syntax annotation (Francom and Hulden, 2008; de la Clergerie et al., 2008). Because of the complexity of syntax annotations as compared to part of speech tags, however, novel technologies have to be applied that allow us to represent, to visualize and to query multiple syntactic analyses of the same sentence.

This paper describes the workflow from raw text to a searchable representation of the corpus. One of the aims of this new resource is to assess potential weaknesses in the parsers as well as their characteristic strengths. For the example of ambiguities in PP attachment, Sect. 4 shows how linguistic analyses can be improved by combining information from different parsers.

2 Parsing

In order to maximize both coverage and granularity of linguistic analyses, we chose parsers from different classes: A probabilistic constituent parser and a rule-based parser that produces semantically enriched dependency parses.

2.1 BitPar

BitPar (Schmid, 2006) is a probabilistic context free parser using bit-vector operations (Schmid, 2004). Node categories are annotated along with grammatical functions, part-of-speech tags and morphological information in a parse tree. BitPar analyses are conformant to the TIGER annotation scheme (Brants et al., 2004), and the tool's output format is similar to the list-based bracketing format of the Penn Treebank (Bies et al., 1995). The BitPar analysis of sentence (1) is visualized as the right-most tree in Fig. 1.

Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pages 166–171,
Uppsala, Sweden, 15-16 July 2010.

(1) *Der Dax reagiert derzeit auf die*
the Dax reacts presently on the
Meldungen aus London.
messages from London
'Presently, the Dax [German stock index, N.B.] is reacting to the news from London.'

2.2 B3 Tool

The second parser applied here is the B3 Tool (Eberle et al., 2008), a rule-based parser that provides syntactic-semantic analyses that combine dependency parsing with FUDRT representations.[1] The B3 Tool is developed on the basis of a research prototype by Lingenio[2] in the context of a project on lexical ambiguities in German nominalizations[3].

For further processing, the output of the B3 Tool is converted into a PTB-style bracketing format similar to that used by BitPar. This transformation involves the generation of a constituency graph from the original dependency analysis: In the first step, rules are used that insert nodes and projections as described by Eberle (2002). Then, another transformation step is necessary: As the B3 Tool aims for an abstract, flat semantics-oriented structure, certain aspects of the surface structure are not represented in its output and need to be restored in order to create analyses that can be aligned with constituent-based representations. For example, punctuation marks do not appear as leaves of the syntactic tree, as their contribution is included in the description of the head verb. Similarly, auxiliaries are not represented as individual words in the B3 output, as their tense and aspect information is integrated with the event description that corresponds to the head verb.[4] As we focus on the integration of multiple syntactic analyses, leaves from the B3 Tool output that represent semantic information were not considered, e.g., information on coreference.

The converted B3 analysis of sentence (1) is visualized as the left tree in Fig. 1.

[1]Flat Underspecified Discourse Representation Theory (Eberle, 1997; Eberle, 2004)

[2]`http://www.lingenio.de/English/`

[3]Project B3 of the Collaborative Research Centre (Sonderforschungsbereich) SFB 732, Stuttgart, Germany.

[4]For the study described here, punctuation marks were added to the surface structure but auxiliaries not yet. There are several possible approaches to dealing with these structural aspects (e.g. inserting empty elements, converting BitPar into B3-like representations, etc.). The discussion of these strategies is, however, beyond the scope of this technical paper.

3 Querying and Visualizing Alternative Parses

In order to integrate multiple annotations created by different tools, we employ a generic XML format, PAULA XML (Dipper and Götze, 2005). PAULA XML is an XML linearization of the data model underlying the ANNIS data base.[5] It is comparable to NITE XML (Carletta et al., 2005) and GrAF (Ide, 2007). PAULA XML supports diverse data structures (trees, graphs, and flat spans of tokens) and allows for conflicting hierarchies.

The integrated PAULA representation of the multiple-parses corpus can be accessed using ANNIS, a web interface for querying and visualizing richly annotated corpora. Fig. 1 shows the ANNIS interface: top left is the query field; below that is the 'match count' field (presenting the number of instances matching the query). Below this field is the list of corpora the user choses from. Matches are visualized in the right window. Tokens and token-level annotations are shown in a Key Word In Context (KWIC) view (upper part of the search result pane in Fig. 1), e.g., B3 morphology (2nd row), BitPar parts of speech (3rd row), and BitPar morphology (4th row). Trees are visualized with the Tree view (below KWIC view).

4 Exploiting multiple parses

The goal of our research is to develop rules for the combination of BitPar and B3 parses such that the resulting merged parse provides more reliable linguistic analyses than the ones provided by either alone. The rule-based B3 Tool provides deep semantic analyses. B3 parses are thus generally richer in information than BitPar parses. Certain ambiguities, however, are not resolved but rather represented by underspecification. In this section, we explore the possibility to employ BitPar parses to resolve such underspecifications.

4.1 Studying PP attachment in ANNIS

The attachment of prepositional phrases is often ambiguous between high attachment (e.g., PP as a clausal adjunct) and low attachment (PP as a nominal modifier). In such cases, the B3 Tool employs underspecification, which is represented by a special edge label `xprep`.[6]

[5]PAULA and ANNIS have been developed at the Collaborative Research Centre 632, `http://www.sfb632.uni-potsdam.de/~d1/annis/`.

[6]The `xprep` label indicates underspecification as to whether the PP has to be attached to its parent node or a node

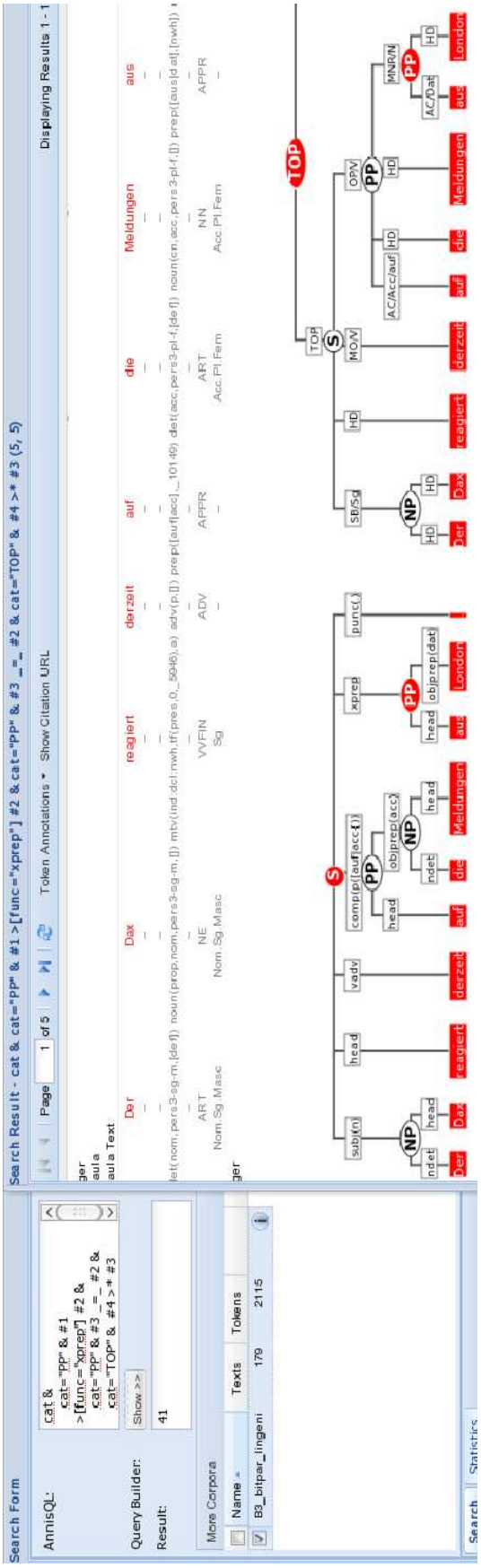

Figure 1: ANNIS2 screenshot with query results for QUERY 1

Using ANNIS, we retrieve all cases where a BitPar PP corresponds to a B3 PP with the edge labeled `xprep` (the query used to accomplish this will be referenced by QUERY 1 in the following). Fig. 1 illustrates an example match: The B3 PP (left tree) is attached to the root node with an edge label `xprep`; in the BitPar analysis (right tree), the prepositional phrase is correctly attached to the other `PP` node.

Using an extended query, we conducted a quantitative analysis comparing the node labels assigned to the parent node of the respective PPs in BitPar parses and B3 parses.

Considering only those matches where the B3 parent node was either VP or S (85%, 35 of 41), high attachment is indicated by BitPar labels VP or S for the BitPar parent node (34%, 12 of 35) and low attachment by labels PP or NP (66%, 23 of 35). BitPar thus distinguishes low and high PP attachment, with a preference for low attachment in our data set.

Results of a subsequent qualitative analysis of the first 20 matches retrieved by this query are summarized in Tab. 1: Only 16% (3 of 19) BitPar predictions are incorrect, 32% (6 of 19) are possible (but different attachment would have produced a felicitous reading), and 53% (10 of 19) are correct. BitPar analyses of PP attachment are thus

BitPar prediction	correct	possible	incorrect	total
low	57%	36%	7%	14
high	40%	20%	40%	5
low or high	53%	32%	16%	19*

* one match (non-sentence) excluded

Table 1: Qualitative analysis of the first 20 matches

relatively reliable, and where the B3 Tool indicates underspecification with respect to PP attachment, the point of attachment can be adopted from the BitPar parse. With such a merging of BitPar parses and B3 parses, a more detailed and more reliable analysis is possible.

4.2 Merging B3 and BitPar parses

With the information from the comparison of BitPar and B3 Tool attachments, a workflow is imaginable where both parsers are applied in parallel, and then their output is merged into a common representation. As opposed to traditional approaches that reduce parse integration to a selec-

dominated by its parent.

tion between entire parses, cf. Crysmann et al. (2002), we employ a full merging between B3 parses and BitPar parses. This merging is based on hand-crafted rules that express preferences between pieces of information from one parse or the other in accordance with the results of quantitative and qualitative analyses as described above.

B3 parses can be enriched with structural information from BitPar, e.g., by the following exemplaric rule:[7] if the B3 parse indicates underspecification with respect to the PP attachment point (QUERY 1), establish a dominance edge between (i) the correspondent of the Bitpar PP (the PP '*from London*' in the example) and (ii) the correspondent of its parent node (the PP '*to the news*'), and delete the original, underspecified B3 edge. The same procedure can also be applied to perform corrections of a parse, if further quantitative and qualitative studies indicate that, for example, the B3 parser systematically fails at a particular phenomenon.

In some cases, we may also want to employ context-dependent rules to exploit the advantageous characteristics of a specific parser, e.g., to preserve ambiguities. Example (2) illustrates that PP attachment has an effect on the sortal interpretation of *Absperrung* 'barrier/blocking/blockade': Different points of attachment can produce different possible readings. The PP *by the police* specifies the subject of the nominalized verb *absperren* 'to block'. This indicates that here, the event/state readings are preferred over the object (=entity) reading.

(2) *Die Feuerwehr unterstützte die*
the fire brigade supported the
Absperrung durch die Polizei.
blocking by the police
'The fire brigade supported the police's blockade/blocking.'

5 Conclusion

In this paper, we described the creation of a resource of German sentences with parallel parses and the infrastructure employed to exploit this resource. We also identified possible fields of application for this resource: By querying this resource one finds strong tendencies regarding the relative reliability and level of detail of different parsers; on this basis, the strengths of several tools can be weighted, as represented, e.g., by generalized, context-dependent rules to combine the output of multiple parsers. Here, this approach was illustrated for two parsers and their combination to disambiguate PP attachment as part of a study of German *-ung* nominalizations. A future perspective could be to add more tools to the comparison, find out their characteristic strengths and perform a sort of weighted voting to decide when an analysis should be enhanced by the information from another one.

We have shown that the infrastructure provided by the ANNIS data base and the underlying data format PAULA can be employed to conduct this kind of research. Although originally developed for different purposes (representation and querying of richly annotated corpora), its generic character allowed us to apply it with more than satisfactory results to a new scenario.

Subsequent research may further exploit the potential of the ANNIS/PAULA infrastructure and the development of application-specific extensions. In particular, it is possible to register in ANNIS a problem-specific visualization for parallel parses that applies in place of the generic tree/DAG view for the namespaces `bitpar` and `b3`. Another extension pertains to the handling of conflicting tokenizations: The algorithm described by Chiarcos et al. (2009) is sufficiently generic to be applied to any PAULA project, but it may be extended to account for B3-specific deletions (Sect. 2.2). Further, ANNIS supports an annotation enrichment cycle: Matches are exported as WEKA tables, statistical, symbolic or neural classifiers can be trained on or applied to this data, and the modified match table can be reintegrated with the original corpus. This allows, for example, to learn an automatic mapping between B3 and BitPar annotations.

Acknowledgements

Collaborative Research Centre 732 (Universität Stuttgart) and Collaborative Research Centre 632 (Humboldt Universität zu Berlin and Universität Potsdam) are funded by Deutsche Forschungsgemeinschaft (DFG).

[7] Other formulations are possible, see Heid et al. (2009) for the enrichment of BitPar parses with lexical knowledge from B3 parses.

References

Marco Baroni and Adam Kilgarriff. 2006. Large linguistically-processed Web corpora for multiple languages. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 87–90, Trento, Italy. EACL.

Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. Bracketing guidelines for treebank ii style penn treebank project. `ftp://ftp.cis.upenn.edu/pub/treebank/doc/manual/root.ps.gz` (May 31, 2010). version of January 1995.

Lars Borin. 2000. Something borrowed, something blue: Rule-based combination of POS taggers. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, May, 31st – June, 2nd.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Jean Carletta, Stefan Evert, Ulrich Heid, and Jonathan Kilgour. 2005. The NITE XML Toolkit: data model and query. *Language Resources and Evaluation Journal (LREJ)*, 39(4):313–334.

Christian Chiarcos, Julia Ritz, and Manfred Stede. 2009. By all these lovely tokens...: merging conflicting tokenizations. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 35–43. Association for Computational Linguistics.

Berthold Crysmann, Anette Frank, Kiefer Bernd, Stefan Mueller, Guenter Neumann, Jakub Piskorski, Ulrich Schaefer, Melanie Siegel, Hans Uszkoreit, Feiyu Xu, Markus Becker, and Hans-Ulrich Krieger. 2002. An integrated architecture for shallow and deep processing. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 441–448, Philadelphia, Pennsylvania, USA, July.

Eric Villemonte de la Clergerie, Olivier Hamon, Djamel Mostefa, Christelle Ayache, Patrick Paroubek, and Anne Vilnat. 2008. PASSAGE: from French Parser Evaluation to Large Sized Treebank. In *Proceedings of the 6^{th} Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May.

Stefanie Dipper and Michael Götze. 2005. Accessing Heterogeneous Linguistic Data — Generic XML-based Representation and Flexible Visualization. In *Proceedings of the 2nd Language & Technology Conference 2005*, pages 23–30, Poznan, Poland, April.

Kurt Eberle, Ulrich Heid, Manuel Kountz, and Kerstin Eckart. 2008. A tool for corpus analysis using partial disambiguation and bootstrapping of the lexicon. In Angelika Storrer, Alexander Geyken, Alexander Siebert, and Kay-Michael Würzner, editors, *Text Resources and Lexical Knowledge – Selected Papers from the 9^{th} Conference on Natural Language Processing (KONVENS 2008)*, pages 145–158, Berlin, Germany. Mouton de Gruyter.

Kurt Eberle, Gertrud Faaß, and Ulrich Heid. 2009. Proposition oder Temporalangabe? Disambiguierung von -ung-Nominalisierungen von verba dicendi in nach-PPs. In Christian Chiarcos, Richard Eckart de Castilho, and Manfred Stede, editors, *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, pages 81–91, Tübingen. Gunter Narr Verlag.

Kurt Eberle. 1997. Flat underspecified representation and its meaning for a fragment of German. Arbeitspapiere des Sonderforschungsbereichs 340, Nr. 120, Universität Stuttgart, Stuttgart, Germany.

Kurt Eberle. 2002. Tense and Aspect Information in a FUDR-based German French Machine Translation System. In Hans Kamp and Uwe Reyle, editors, *How we say WHEN it happens. Contributions to the theory of temporal reference in natural language*, pages 97–148. Niemeyer, Tübingen. Ling. Arbeiten, Band 455.

Kurt Eberle. 2004. Flat underspecified representation and its meaning for a fragment of German. Habilitationsschrift, Universität Stuttgart, Stuttgart, Germany.

Gertrud Faaß, Ulrich Heid, and Helmut Schmid. 2010. Design and application of a Gold Standard for morphological analysis: SMOR as an example of morphological evaluation. In *Proceedings of the seventh international conference on Language Resources and Evaluation (LREC)*, Valetta, Malta.

Jerid Francom and Mans Hulden. 2008. Parallel Multi-Theory Annotations of Syntactic Structure. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May.

Ulrich Heid, Kurt Eberle, and Kerstin Eckart. 2009. Towards more reliable linguistic analyses: workflow and infrastructure. Poster presentation at the GSCL 2009 workshop: Linguistic Processing Pipelines, Potsdam.

Nancy Ide. 2007. GrAF: A Graph-based Format for Linguistic Annotations. In *Proceedings of the LAW Workshop at ACL 2007*, Prague.

Helmut Schmid. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proceedings of the 20th International Conference on Computational Linguistics, Coling'04*, volume 1, pages 162–168, Geneva, Switzerland.

Helmut Schmid. 2006. Trace Prediction and Recovery With Unlexicalized PCFGs and Slash Features. In *Proceedings of COLING-ACL 2006*, Sydney, Australia.

Dan Tufiş. 2000. Using a large set of EAGLES-compliant morpho-syntactic descriptors as a tagset for probabilistic tagging. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pages 1105–1112, Athens, Greece, May, 31st – June, 2nd.

Jakub Zavrel and Walter Daelemans. 2000. Bootstrapping a Tagged Corpus through Combination of Existing Heterogeneous Taggers. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, May, 31st – June, 2nd.