# Evaluating Dependency Parsing Performance on German Learner Language

**Niels Ott and Ramon Ziai**

Collaborative Research Center 833, Project A4
University of Tübingen
E-mail: {nott,rziai}@sfs.uni-tuebingen.de

**Abstract**

We present an experiment on dependency parsing of German learner language. Ultimately aiming at evaluating the meaning of learner answers to German reading comprehension questions, we are interested in how reliable a parser trained on native language can identify the main argument relations. To that end, we manually annotated a small set of learner answers and parsed it using MaltParser (Nivre et al., 2007) trained on TüBa-D/Z (Telljohann et al., 2004). The evaluation of the results shows that semantically salient relations such as SUBJ and OBJ can generally be found reliably. Qualitative analysis indicates that the omission of syntactically central material, such as the finite verb, yields incorrect parses while other errors, e.g. in agreement or word order, can still be parsed robustly.

## 1 Introduction

In this paper, we present a pilot study about dependency-parsing German learner language with MaltParser (Nivre et al., 2007) trained on the TüBa-D/Z treebank (Telljohann et al., 2004). The context of this pilot study is an on-going research project on meaning comparison in realistic situations. Building on Bailey & Meurers (2008), we are exploring ways of evaluating student answers in reading comprehension tasks with respect to both the reading comprehension questions and the target answers given by language teachers.

Automatic dependency analysis of learner language is one component out of many in such a content assessment system. In this paper, we focus on this type of analysis as a separate subject of investigation, yet with an emphasis on our research intentions in the overall project. Unlike other projects involving learner language, we are not investigating form errors or L2 development. Our interest lies in the machinery required to perform robust analyses, supporting the creation of semantic representations, on several levels of complexity, given that the input often is not well-formed language.

For the pilot study, we worked on a snapshot of 106 learner answers from a task-based learner corpus which is currently being compiled within the project (Meurers, Ott & Ziai, 2010). This snapshot was annotated by three independent annotators using the dependency grammar scheme devised by Foth (2006). The procedure of selecting and annotating data as well as the peculiarities of annotating learner language are described in detail in section 3 after section 2 presents related work.

Since the corpus used in this study is not very large, we do not only look at quantitative evaluation measures: section 4 deals with the parsing procedure and also contains a qualitative analysis of selected issues arising in the automatic dependency analysis of learner language. We also take a look at automatic part-of-speech tagging and its influence on parser performance.

## 2 Related Work

Parsing learner language is not a novelty. Most notably, learner language is automatically analyzed in Intelligent Language Tutoring Systems, such as *e-tutor* (Heift & Nicholson, 2001), *Robo-Sensei* (Nagata, 2009) and *TAGARELA* (Amaral, 2007; Amaral & Meurers, 2011). Here, parsing is employed with the ultimate aim of giving feedback to students, mainly on form errors. Therefore, the parsing strategy needs to account for and explicitly model learner errors in a way that allows error detection and feedback (see e.g. Menzel & Schröder 1999).

More recently, there is some research on devising syntactic annotation schemes specific to language acquisition in progress, so-called *interlanguage*. Dickinson & Ragheb (2009) present such an approach, analyzing and annotating the interlanguage of English language learners. Most closely related to our work is the experiment presented by Dickinson & Lee (2009), where corpus annotation has been altered to support training of a dependency parser capable of handling a limited range of learner errors, namely postpositional particles in Korean.

However, in this paper, we are not concerned with giving feedback to learners or accurately describing interlanguage. Instead, we want to robustly parse learner language in order to access its content, and subsequently compare such content against that of reference answers. Therefore, due to our different motivation, we instead investigate how a parser trained on native language behaves when confronted with learner language. To our knowledge, such an experiment has not been done for German with a hand-annotated gold standard.

## 3 Annotation of Learner Answers

In this section, we describe the creation of a small corpus of learner answers to reading comprehension questions and its manual annotation with dependency grammar analyses. This small corpus serves as the gold standard for parser evaluation in the experiment described in the presented paper.

### 3.1 Data Source

The data used in the presented experiment and in our research project is collected in large German programs in the US, at Kansas University (Prof. Nina Vyatkina) and The Ohio State University (Prof. Kathryn Corl). Using the WEb-based Learner COrpus MachinE (WELCOME), a tool that has been developed especially for this purpose, German teachers in the two programs are collecting reading comprehension exercises consisting of texts, questions, target answers, and corresponding student answers (Meurers, Ott & Ziai, 2010). Each student answer is transcribed from the hand-written submission by two independent annotators. These two annotators also assess the contents of the answers purely on the basis of meaning: Did the student answer convey the meaning that was required by the question?

The corpus emerging from our on-going four-year collection phase is steadily growing. For this paper, we took a snapshot and selected learner answers according to the following criteria: a) full agreement in meaning assessment, b) edit distance between the two transcriptions of handwriting at most 1, c) minimum length of five tokens. Inevitably, answers by different students to one and the same reading comprehension question are quite similar. In order to ensure variety in our data, we randomly selected only one student answer for each question after applying the constraints a)–c).

The resulting subcorpus consists of 106 learner answers containing 109 sentences from the beginner and intermediate levels of the respective German programs. The average sentence length is 8.26 tokens with a standard deviation of 3.11. The shortest sentence contains 2 tokens, the longest 17. Tokenization and sentence segmentation were performed automatically using the OpenNLP[1] components and their default statistical models.

### 3.2 The Annotation Process

As far as the choice of dependency annotation scheme is concerned, we first looked at the ones employed at the CoNLL-X shared task on dependency parsing (Buchholz & Marsi, 2006). This task used a version of the TIGER treebank (Brants et al., 2002) converted to dependency grammar. However, we quickly found it difficult to annotate our data using this scheme because it is based on the phrase structure backbone of the TIGER annotation scheme (Albert et al., 2003). Thus, constructing a dependency analysis with the TIGER scheme manually would have required us to produce a phrase structure-based analysis first, which was not our intention. In contrast, Foth (2006) provides a readily available scheme and manual for German dependency grammar that we found convenient to use for the task of manual annotation.

Consequently, we used a dependency grammar version of the Tüba-D/Z treebank (Telljohann et al., 2004), because using Versley (2005)'s approach, it can be converted to the scheme devised by Foth (2006). However, relying only on

[1] `http://opennlp.sourceforge.net`

Foth's dependency grammar model in the annotation process would have mixed two variables in the parser evaluation step: the difference between Foth (2006) and the auto-converted Tüba-D/Z annotation (Versley, 2005), and the actual performance of the parser trained on Tüba-D/Z but parsing learner language. To minimize this problematic effect, annotators were advised to use the auto-converted Tüba-D/Z as the definite resource in case Foth (2006) was unclear or incomplete.[2]

The 109 sentences of our subcorpus were annotated by three independent annotators using DgAnnotator[3]. The percentage agreement between the three annotators computed to 88.1% for labeled attachment. However, we found only 67 sentences for which at least two annotators had fully agreed on the dependency annotation (both relation labeling and attachment). For the remaining 42 sentences, two of the annotators independently examined the differences in the annotations, writing down comments on the most salient issues having caused the deviation in the annotations. The third annotator then served as a judge fixing the annotations of these 42 sentences according to the comments.

Part-of-speech (POS) annotation was conducted automatically using Tree Tagger (Schmid, 1994) equipped with the standard model for German. This standard model uses the Stuttgart Tübingen Tagset (STTS, Thielen et al. 1999). Since the STTS variant used in Tüba-D/Z is slightly different, we converted the tagger output to fit the POS annotation that the parser had seen during training. Annotators were advised to manually correct the annotation on the POS layer in the case of tagging errors. The POS layer was double-checked by the abovementioned third annotator for correctness and consistency. As a result of this step, 7.2% of all POS labels in our subcorpus were manually post-corrected.

The context of the reading comprehension exercises consisting of reading texts, questions, and target answers was not taken into account in the annotation process. Similarly to the parser, the annotators did not have any additional sources of context knowledge for interpreting the learner answers.

### 3.3 Specific Issues in Annotating Learner Language

A popular theme in second language acquisition (SLA) research is error tagging of learner corpora (see e.g. Granger et al. 2009). Error tagging refers to the process of annotating defects in learner language with regard to well-formed versions of utterances, also called target hypotheses (Lüdeling et al., 2005). Target hypothesis are highly subjective: Lüdeling (2008) found that in an essay written by an advanced learner of German, there was not a single one out of 17 sentences for which all of her five professionally trained German teachers agreed on one single target hypothesis.

Dickinson & Ragheb (2009) propose a scheme for describing the learner's interlanguage instead of the L2 that the learner is about to acquire. In contrast, we admittedly use native language categories for annotating learner language. There are several reasons for following this route: first, a practical reason is that annotated

[2]More precisely speaking, annotators used only the test set as a resource. See also section 4.

[3]`http://medialab.di.unipi.it/Project/QA/Parser/DgAnnotator/`

*Learner sentence:*

'Sie dachte es, welche das schönste Puppenwagen, den sie je gesehen hatte.'

*Target hypothesis 1:*

Sie dachte es, welcher der schönste Puppenwagen [ist], den sie je gesehen hatte
She thought it, which the finest doll's pram [is] that she ever seen had

*Target hypothesis 2:*

Sie dachte, es wäre der schönste Puppenwagen, den sie je gesehen hatte
She thought, it was the finest doll's pram, that she ever seen had

Figure 1: Differing target hypotheses due to learner errors.

corpora for training a parser need to be large and they are only available for native language at the moment. Second, for the purposes of our project we need to compare learner answers with target answers, questions, and reading texts written by native or near-native speakers. However, an interlanguage category system would not necessarily be applicable to native text. And third, interlanguages are specific to both the learner's background and his stage of acquisition, and thus inherently difficult to capture in a single annotation scheme or parsing model.

For a number of sentences in our small corpus, we found it impossible to annotate them without constructing a target hypothesis. Among the 42 problematic sentences mentioned earlier, there were six cases in which the two commenters agreeingly attributed the deviation in the dependency annotation to differing target hypotheses. One such case is presented in Figure 1, where several learner errors make interpretation difficult.

For further research we therefore propose to follow the route suggested by Lüdeling (2008): target hypotheses should be explicitly annotated in the corpus. However, since we do not make use of them in our automatic dependency analysis, other users of the data would be the ones to benefit more from such an annotation.

# 4 Parsing Results

## 4.1 Setup

We used MaltParser (Nivre et al., 2007) for our parsing experiments. As previously mentioned, we applied the conversion procedure presented by Versley (2005) to release 5 of the TüBa-D/Z treebank (Telljohann et al., 2004) in order to get a training data basis that uses the Foth (2006) annotation scheme. We used 90% (40676 sentences) of the TüBa-D/Z sentences for training the parser, leaving out 10% (4520 sentences) to be used as a native language evaluation set, a point of reference for our experiments with learner language. The parameters used for training MaltParser where taken from Gómez-Rodríguez & Nivre (2010), using MaltParser's new 2-planar algorithm which is capable of handling some non-projectivity. The resulting

| | Learner LAS | TüBa LAS | Learner UAS | TüBa UAS |
|---|---|---|---|---|
| Automatic POS | 79.15% | 83.12% | 84.81% | 86.38% |
| Manual POS | **85.71%** | 88.07% | **90.22%** | 90.50% |

Table 1: Overall attachment scores for learner data set and TüBa-D/Z test set (LAS: labeled attachment score, UAS: unlabeled attachment score).

model achieves 88.07% labeled attachment score (LAS) on our TüBa-D/Z test set, which is state-of-the art.

Using this model, we parsed the hand-annotated learner data set presented in section 3. In order to be able to assess the contribution of proper POS annotation, we ran the parser on both automatically tagged and manually corrected versions of the data set. However, since a performance difference could also be due to the fact that automatic POS tagging is simply always inferior in quality to manual inspection, and not a result of learner language specifics, we made the same distinction with the TüBa-D/Z test set.

## 4.2 Quantitative Evaluation

The quantitative results are summarized in Table 1. All figures were obtained using the *eval.pl* script from the CoNLL-X shared task on dependency parsing (Buchholz & Marsi, 2006), which does not depend on any particular dependency scheme. Naturally, quantitative results on a small data set like the one we annotated should be interpreted with some caution as far as representativeness is concerned. However, it seems that the parsing results on our learner data set are within acceptable range ($\approx 3\%$) of what is currently considered to be state-of-the art in dependency parsing. Interestingly, automatic POS annotation causes a similar performance drop in both data sets. This indicates that there is no clear difference in the parsing contribution of higher quality POS between learner and native language. Note also the relatively high difference between labeled and unlabeled attachment scores for the learner data set, suggesting that the gap between knowing where to attach a word and how to label the attachment is wider for learner language, which seems plausible given the ungrammatical constructions learners sometimes use.

We also looked at parsing performance on individual dependency relation types. Recall that our ultimate goal is to obtain a representation of learner utterances suitable for semantic comparison with a reference answer. For such a representation, functor-argument relations such as `SUBJ` and `OBJA` thus seem particularly important, because they specify the main verb's arguments. Following Foth (2006, p. 6)'s distinction of argument from modifier relations, we summarized the ones which were interesting to us and had more than 10 gold standard instances in Table 2. The figures represent combined scores for both correct dependency labels and correct attachments. (An overview of the dependency labels used in this paper is given in Table 3 in the appendix.)

| | Learner Data Set | | | TüBa-D/Z Test Set | | |
|---|---|---|---|---|---|---|
| | Recall | Precision | Count | Recall | Precision | Count |
| | Argument relations | | | | | |
| DET | 99.02% | 97.12% | 102 | 99.15% | 99.51% | 9272 |
| SUBJ | 90.53% | 88.66% | 95 | 88.10% | 91.37% | 6043 |
| OBJA | 84.31% | 84.31% | 51 | 81.77% | 77.70% | 2979 |
| CJ | 87.18% | 89.47% | 39 | 93.91% | 94.03% | 2514 |
| PRED | 65.38% | 85.00% | 26 | 75.80% | 80.38% | 1157 |
| AUX | 100.00% | 95.83% | 23 | 93.93% | 97.03% | 2503 |
| OBJP | 36.36% | 80.00% | 11 | 55.92% | 69.13% | 769 |
| | Modifier relations | | | | | |
| KON | 63.27% | 65.96% | 49 | 81.22% | 79.09% | 3003 |
| ADV | 68.18% | 73.17% | 44 | 83.76% | 84.38% | 5770 |
| PP | 80.00% | 55.81% | 30 | 75.25% | 75.07% | 6302 |

Table 2: Precision and recall for selected dependency types including attachment. The figures are based on the parser output based on gold POS tags and the *Count* columns represent the number of occurrences in the respective gold standards.

According to the figures we obtained, subject and (accusative) object relations can be found approximately as reliably as in native language text, with precision and recall around 90% for subjects and 84% for accusative objects. Predicates are missed more often, with a recall of just 65.38%, but tend to be reliable if identified. Prepositional complements are often analyzed as adjuncts, indicated by the low recall for `OBJP` (prepositional object) and relatively low precision for `PP` (prepositional adjunct). However, this distinction was also difficult for human annotators, since it is not always obvious whether a prepositional element is an obligatory verb argument or not.

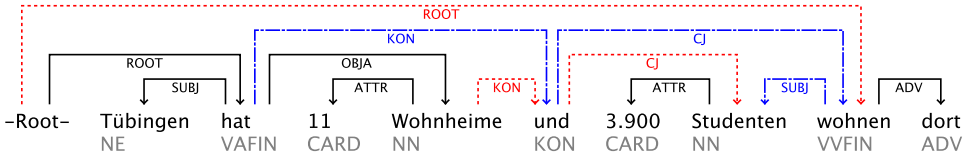
### 4.3 Qualitative Evaluation

We also qualitatively evaluated our parsing results, partly because our data set is small and thus quantitative means of evaluation are of limited usefulness, and also because it is instructive to closely inspect the problems a parser encounters when confronted with learner data. In the following figures[4], dash-dotted arcs represent the gold standard whereas dotted red arcs represent the parser's decisions. Solid lined arcs represent correctly analyzed structures. Where necessary for explanation, we included target hypotheses although as mentioned in section 3, we did not explicitly construct these during annotation.

[4]We used *What's wrong with my NLP?* for generating the figures, see `http://code.google.com/p/whatswrong`

### Common Dependency Parsing Problems

Unsurprisingly, some of the issues are already well-known from parsing native language. Most notably, coordination is problematic and occurs quite frequently which, given the prompts for lists of items or propositions students are faced with, is to be expected. Consider the sentence in Figure 2.




'Tübingen has 11 dormitories and 3.900 students live there'

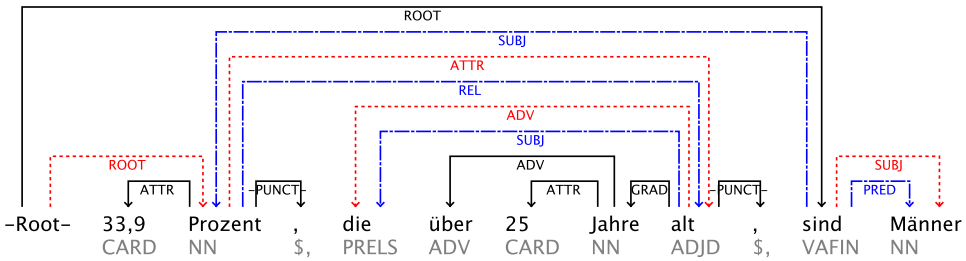
Figure 2: Sentence-level coordination wrongly analyzed.

The sentence is a perfectly well-formed example of sentence-level coordination, however the parser seems to have problems attaching conjuncts over longer distances. This might be due to the fact that the feature models underlying most parsers, including the one used here, only look a few tokens behind or ahead.

### Learner Errors Not Handled Well

Of course there are also examples where the very nature of learner language made a correct dependency analysis in terms of L2 impossible. Usually these were the cases where also the human annotators were unsure of how to annotate a given sentence. Consider the sentence in Figure 3, where the learner left out the finite verb of a relative clause.

In the given target hypothesis of this sentence, the finite verb in the relative clause would have been the link to the main clause. However, since that verb is missing, the parser used the adjective *alt* ('old') for this purpose. Incidentally, so did the human annotators, albeit with different labels for the relations involving *alt*.




*Target hypothesis:*

33,9 Prozent, die über 25 Jahre alt [sind], sind Männer.
33.9 percent, who over 25 years old [are], are men.

Figure 3: Dependency errors due to missing verb in relative clause.

**Learner Errors Handled Well**

On the bright side, some learner errors also seem to be handled well. Specifically, local problems such as wrong verb forms or word order are usually handled correctly. The sentence in Figure 4 is such an example.

Here, the learner apparently overused participles, namely *gekramt* instead of preterite *kramte* ('rummaged') and *aufgefunden* instead of infinitive *finden* ('to find'). Moreover, *wollte* ('wanted') ought to be at the end of the subordinate clause, not next to the subject *sie* ('she'). However, all of these errors are analyzed robustly and the parser comes up with the analysis proposed by the human annotators.

# 5 Conclusion

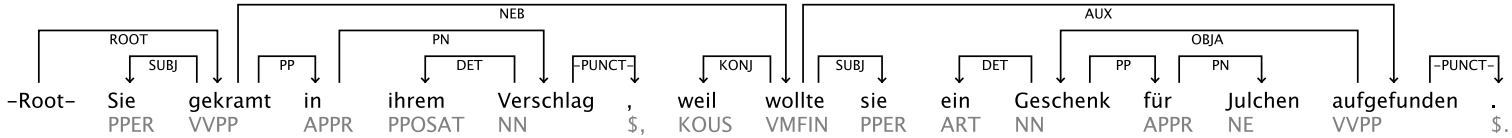
We presented a parsing experiment on German learner language using a state-of-the art dependency parser trained on native language. In order to evaluate the result, we hand-annotated a sample of the German learner corpus we are currently collecting. The results are generally promising, showing that the main functor-argument relation types we are most interested in can generally be identified reliably, with precision and recall in the area of 80–90%. Furthermore, qualitative evaluation shows that while learner errors that result in the omission of syntactically central material lead to serious problems in dependency analysis, other errors are often handled well. The latter include agreement errors, verb tense errors and word order errors.

Future work should include the annotation of more data from several levels of difficulty in order to obtain a wider range of linguistic phenomena, making a quantitative evaluation more meaningful. An interesting addition to explore would be to also annotate the learner sentences with well-formedness judgments, enabling us to find out whether there is a significant correlation between ill-formedness and parsing errors. The annotation of target hypotheses could make the implicit assumptions underlying the dependency analysis explicit in the future, which would facilitate the interpretation of the annotation.

Concerning our intent of evaluating the meaning of learner answers, we are now more confident that a dependency analysis of German learner language based on an L2 parser model is a viable method, provided one is "careful" with the interpretation of the parse and knows what kind of input will potentially result in bad analyses such as the one in Figure 3. Going forward, we are planning to investigate methods of transforming dependency trees into semantically more useful representations, possibly in the spirit of the 'compositional facets' employed by Nielsen et al. (2009) in the context of Intelligent Tutoring Systems.

## Acknowledgements

We wish to thank Detmar Meurers and Holger Wunsch for their valuable comments. We are also thankful to Janina Kopp who was of great assistance in the annotation process. Finally, we also benefited from three anonymous workshop reviews.




Target hypothesis:

Sie kramte in ihrem Verschlag, weil sie ein Geschenk für Julchen finden wollte.
She rummaged in her shed, because she a gift for Julchen to-find wanted.

'She rummaged in her shed because she wanted to find a gift for Julchen.'

Figure 4: Robust analysis of ill-formed sentence.

## References


Albert, S., J. Anderssen et al. (2003). *TIGER Annotationsschema*. Universität des Saarlandes and Universität Stuttgart and Universität Potsdam.

Amaral, L. (2007). Designing Intelligent Language Tutoring Systems: integrating Natural Language Processing technology into foreign language teaching. Ph.D. thesis, The Ohio State University.

Amaral, L. & D. Meurers (2011). On Using Intelligent Computer-Assisted Language Learning in Real-Life Foreign Language Teaching and Learning. *ReCALL* 23(1).

Bailey, S. & D. Meurers (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In J. Tetreault, J. Burstein & R. D. Felice (eds.), *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*. Columbus, Ohio, pp. 107–115.

Brants, S., S. Dipper, S. Hansen, W. Lezius & G. Smith (2002). The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*. Sozopol.

Buchholz, S. & E. Marsi (2006). CoNLL-X shared task on multilingual dependency parsing. In *CoNLL-X '06: Proceedings of the Tenth Conference on Computational Natural Language Learning*. Morristown, NJ, USA: Association for Computational Linguistics, pp. 149–164.

Dickinson, M. & C. M. Lee (2009). Modifying Corpus Annotation to Support the Analysis of Learner Language. *CALICO Journal* 26(3), 545–561.

Dickinson, M. & M. Ragheb (2009). Dependency Annotation for Learner Corpora. In *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT-8)*. Milan, Italy.

Foth, K. (2006). *Eine umfassende Constraint-Dependenz-Grammatik des Deutschen*. Tech. rep., Universität Hamburg.

Gómez-Rodríguez, C. & J. Nivre (2010). A Transition-Based Parser for 2-Planar Dependency Structures. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. pp. 1492–1501.

Granger, S., E. Dagneaux, F. Meunier & M. Paquot (2009). *International Corpus of Learner English Version 2*. Presses Universitaires de Louvain.

Heift, T. & D. Nicholson (2001). Web Delivery of Adaptive and Interactive Language Tutoring. *International Journal of Artificial Intelligence in Education* 12(4), 310–325.

Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In P. Grommes & M. Walter (eds.), *Fortgeschrittene Lernervarietäten*, Tübingen: Niemeyer, pp. 119–140.

Lüdeling, A., M. Walter, E. Kroymann & P. Adolphs (2005). Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics*. Birmingham.

Menzel, W. & I. Schröder (1999). Error diagnosis for language learning systems.

*ReCALL* Special ed., 20–30.

Meurers, D., N. Ott & R. Ziai (2010). Compiling a Task-Based Corpus for the Analysis of Learner Language in Context. In *Proceedings of Linguistic Evidence*. Tübingen, pp. 214–217.

Nagata, N. (2009). Robo-Sensei's NLP-Based Error Detection and Feedback Generation. *CALICO Journal* 26(3), 562–579.

Nielsen, R. D., W. Ward & J. H. Martin (2009). Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering* 15(4), 479–501.

Nivre, J., J. Nilsson, J. Hall, A. Chanev, G. Eryigit, S. Kübler, S. Marinov & E. Marsi (2007). MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. *Natural Language Engineering* 13(1), 1–41.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK, pp. 44–49.

Telljohann, H., E. Hinrichs & S. Kübler (2004). The TüBa-D/Z Treebank: Annotating German with a Context-Free Backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. Lissabon.

Thielen, C., A. Schiller, S. Teufel & C. Stöckert (1999). *Guidelines für das Tagging deutscher Textkorpora mit STTS*. Tech. rep., Institut für Maschinelle Sprachverarbeitung Stuttgart and Seminar für Sprachwissenschaft Tübingen.

Versley, Y. (2005). Parser Evaluation across Text Types. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT)*. Barcelona, Spain.

## Dependency Labels

| **Label** | **Short description** | **Label** | **Short description** |
|---|---|---|---|
| ADV | adverbial modifier | OBJA | accusative object |
| ATTR | nominal attribute | OBJP | prepositional object |
| AUX | auxiliary verb | PN | preposition complement |
| CJ | complement of a conjunction | PP | prepositional adjunct |
| DET | determiner of a noun | PRED | predicate |
| GRAD | accusative NP as measuring unit | -PUNCT- | punctuation |
| KON | non-final coordination conjunct | REL | relative clause |
| KONJ | subordinating conjunction | ROOT | root of the sentence |
| NEB | subordinate clause | SUBJ | subject |

Table 3: Dependency labels used in this paper, taken from Foth (2006).