Linguistic **I**ssues in **L**anguage **T**echnology – **LiLT**
Submitted, January 2012

Linguistic Annotations for a Diachronic Corpus of German

Erhard Hinrichs and Thomas Zastrow

Published by CSLI Publications

LiLT volume 7, issue 7 January 2012

Linguistic Annotations for a Diachronic Corpus of German

Erhard Hinrichs and Thomas Zastrow, *Seminar für Sprachwissenschaft, Eberhard Karls Universität Tübingen*

Abstract

This paper describes the TüBa-D/DC, a diachronic corpus of German that uses selected materials from the German Gutenberg Project and enriches them with different linguistic annotation layers, including part-of-speech, lemmata, and constituent structure. Linguistic annotation is performed automatically by using statistical tools that have been trained with data from the *Tübinger Baumbank des Deutschen* (TüBa-D/Z). In order to assess the annotation quality, an evaluation of the POS tagging is performed on the basis of a data sample of texts that range from the 13th to the 20th century. The paper concludes with a description of three different query mechanisms provided for the user.

LiLT Volume 7, Issue 7, January 2012.
Linguistic Annotations for a Diachronic Corpus of German.

1 Introduction

The availability of large amounts of linguistically annotated text corpora is essential for data-driven research in computational linguistics, corpus linguistics and empirically grounded theoretical linguistics. Such corpora should ideally satisfy four criteria:

1. They should be freely available – and thus sharable – for academic use.
2. They should include materials of different kinds and text genres.
3. They should be of sufficient size.
4. They should be of sufficient quality, i.e. contain few typos and ideally only textual material proper[1].

In the past, corpus collection efforts which have focused on the criterion of sufficient size and quality have concentrated on synchronic newspaper data. At least in some cases, such materials also fulfilled the criterion of being freely available – especially in the case of treebanks which use only a small portion of a larger newspaper archive.

The present paper focusses on a linguistically annotated resource which promises to fulfill as many of the above mentioned criteria for corpus collection without claiming to satisfy all of them completely. This resource uses selected materials from the *German Gutenberg Project*[2] (henceforth referred to as GGP) and enriches them with different linguistic annotation layers. The selected language materials satisfy the criterion of availability since the chosen Gutenberg materials are copyright-free. They are of sufficient size, contain different genres, and cover many centuries. However, since the origin of the data is often unknown, the materials are not always guaranteed to fulfill the same exacting standards of data quality. The linguistic annotations are generated automatically by a suite of widely used tools from computational linguistics. The suite includes syntactic parsing so that the resulting resource contains treebank data, but also additional layers of annotation, such as named entity information. We will subsequently refer to this resource as the TüBa-D/DC (short for: *Tübinger Baumbank des Deutschen/Diachrones Corpus*).

The remainder of this paper is structured as follows: Section 2 describes the underlying corpus data and the process of automatic annotation in more detail. Section 3 focusses on the details of the annotation,

[1] Here we refer to the known problem of web-harvested materials, but to some extent also newspaper materials, which often contain unwanted tables, links etc.

[2] The Gutenberg Project (http://gutenberg.spiegel.de/) is a community-driven initiative of volunteers, not of professional editors.

including the annotation layers and the underlying data format. Section 4 addresses issues of annotation quality and reports on a manual evaluation of the part-of-speech annotations. Section 5 describes ways of querying the data and the linguistic annotations at different levels of granularity. The paper concludes with a discussion of future directions of research and of planned further corpus development.

2 The Data

Number of authors:	875
Number of texts:	19.377
Number of Tokens:	252.520.365
Number of Sentences:	11.713.512
Time period covered:	1210 - 1930
Text genres (incomplete list):	Short stories, novellas, novels, plays, poetry, letters, fairy tales, autobiography and essays

TABLE 1 Profile of the TüBa-D/DC

As mentioned in the previous section, the materials selected for the TüBa-D/DC are taken from the GGP. The main properties of the TüBa-D/DC are listed in Table 1.

Several deliberate decisions were made in the selection of GGP materials: (i) while the GGP also includes texts in languages other than German, the TüBa-D/DC only contains German texts, (ii) while the GGP also contains materials dated prior to 1210, these materials were excluded because they are often not 'authentic' in the sense that they are High German translations of originally Latin and Classical Greek texts, (iii) while the GGP also includes special purpose texts such as recipes, tabular data (for example, a comparison of the size of competing armies), etc., such texts are not well suited for the purpose of creating treebank data.

It needs to be emphasized that the TüBa-D/DC is not meant as a competing approach to other on-going initiatives that are developing linguistically annotated diachronic corpora for German such as the project *DeutschDiachronDigital*[3] (DDD), which is constructing a German reference corpus that covers German texts from the beginnings of the written tradition to the present time. Rather, the focus of these two projects is complementary: while the DDD focusses on represen-

[3] http://www.deutschdiachrondigital.de/

tative materials with high quality, manually checked annotations for a text collection of a rather modest size, the TüBa-D/DC aims at a much larger text collection with automatically created, and thus noisy annotations. This reflects the usual "size versus reliability" trade-off for linguistically annotated corpora, where resources of one type are no substitute for the other.

The materials selected for the TüBa-D/DC come with a set of metadata (about authorship, publication dates, etc.) contained in the GGP. These metadata are often incomplete or misleading (for example by not documenting the actual textual source that served as the input for the digital edition in the GGP)

We manually checked the metadata for all the 19,377 data files (cf. Table 1) and tried to add information where possible. However, the updated metadata are still far from perfect. Goethe's novel Die Leiden des jungen Werther, one of the texts contained in the GGC, is a good example of the complexities that arise when one tries to specify the actual textual source that served as the input for the digital edition in the GGP. The Werther novel was published in the 18th century in two editions authorized by Goethe himself. The first edition was published in 1774 and replaced by Goethe himself in 1787 by the second edition. This second edition was later incorporated into the *Weimarer Ausgabe* of 1899. The GGP version of the Werther text is a modified version of the *Weimarer Ausgabe* with some orthographical changes to reflect the rules of modern German orthography. This information can be reconstructed with the help of the metadata published in the GGC, albeit only in combination with consulting the published works on which it is based.[4] We have described this example in some detail, so as to show that textual tradition can be highly complex especially for historical texts. The often voiced criticism against projects such as the GGP needs to be taken with a grain of salt because it is sometimes based on the naive and erroneous assumption that there is a unique authoritative text that should have been captured in the GGP digital edition. Alas, philological reality is far more complex and can often not be fully modeled, even in the most fine-grained set of metadata.

Table 2 provides a synopsis of the different levels of annotation included in the TüBa-D/DC. For each level, it identifies the automatic tool that was used to produce the annotation[5].

[4]In the case of Goethe's Werther, the metadata contains the ISBN number of the book from which the digital edition was produced.

[5]The tools can be found at the following URLs:

- OpenNLP project: http://incubator.apache.org/opennlp/
- TreeTagger: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Layer	Tool Used
Tokens	OpenNLP tokenizer
Part of speech tags	TreeTagger
Lemmas	TreeTagger
Sentence boundaries	in-house tool
Named entities (persons, locations, organizations)	in-house tool
Constituent parse trees	Berkeley Parser

TABLE 2 Annotation levels of the TüBa-D/DC

For all tools that require a trained data model, version 5.0 of the German treebank *Tübinger Treebank des Deutschen/Zeitungstext* (Telljohann et al., 2004, 2009), was used as training data.[6] The TüBa-D/Z consists exclusively of contemporary newspaper text. This raises issues of out-of-domain language models since the TüBa-D/DC data are genre diverse and diachronic in character. These issues will be taken up in detail in Section 3 below. Where no external tools were freely available for analysis of German data or would fit the requirements of the task at hand, self-developed tools were used.

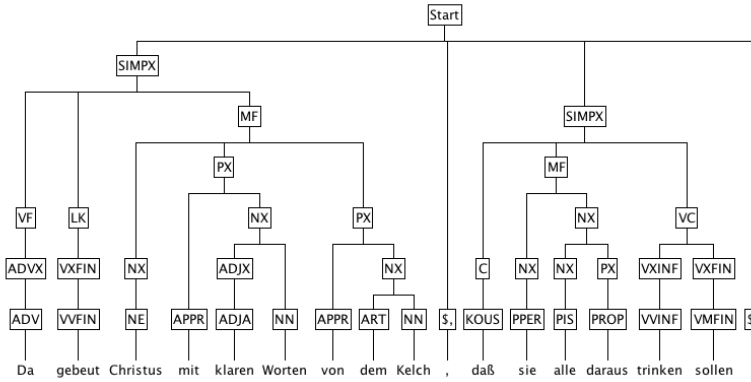

FIGURE 1 A parse tree from the subcorpus Wunderlicher Traum von einem grossen Narrennest

· Berkeley Parser: http://code.google.com/p/berkeleyparser/

[6]This pertains to the OpenNLP tokenizer, the TreeTagger, and the Berkeley Parser.

In order to ensure interoperability of the tools used in the annotation tool chain, a common text format for processing the input data of each layer was a necessary prerequisite. For this purpose, the *Text Corpus Format* (TCF) was used. TCF (Heid et al., 2010) is the processing format orginally designed to be used in *WebLicht* (Hinrichs et al., 2010), a service-oriented architecture for distributed deployment of heterogenous NLP tools.[7]

Figure 1 shows the parse tree of the sentence *Da gebeut Christus mit klaren Worten von dem Kelch, daß sie alle daraus trinken sollen* ('There ordered Christ with clear words about the chalice that they all should drink from it') taken from the text Wunderlicher Traum von einem grossen Narrennest (see Table 3)[8]. The syntactic structure is generated by the Berkeley parser that has been trained on the TüBa-D/Z. The preterminal nodes in the parse tree are labeled by part of speech tags taken from the *Stuttgart Tübingen Tagset* (STTS, Schiller et al. (1995)). Notice that the final verb *gebeut* is correctly tagged, even though it is an archaic word and, thus, an out-of-vocabulary item for the part-of-speech tagger and for the parser, which were both trained on the TüBa-D/Z. The phrasal nodes include a layer of topological fields such as VF (*Vorfeld*), LK (*Linke Klammer*), MF (*Mittelfeld*) and VC (*Verbalkomplex*). Topological fields in the sense of (Höhle, 1986, Herling, 1821, Drach, 1937) are widely used in descriptive studies of German syntax. Such fields constitute an intermediate layer of analysis above the level of individual phrases and below the clause level.

3 Data Evaluation

Since the linguistic annotation of the TüBa-D/DC was performed automatically, the resulting annotations will not be 100% accurate. The fact that the part-of-speech tagger and the constituent parser use out-of-domain language models raises additional concerns about annotation qualities. It is well-known that training of statistical models for taggers and parsers on out-of-domain data leads to a significant drop in accuracy already for synchronic data from different domains (Bikel, 2004, Kübler and Baucom, 2011), and is even more problematic when

[7] TCF is fully compatible with existing relevant ISO standards such as LAF or MAF (for more details, see Heid et al. 2010). Conversion from various formats in and out of TCF can be easily performed with the help of existing conversion tools provided in the WebLicht environment.

[8] The parse tree was created with *TViewer*, an application for visualizing the linguistic layers of a TCF file. See http://de.clarin.eu/index.php/weblicht/tutorials/57-tcf/30 for more information.

such models are applied to heterogeneous diachronic materials (Dipper, 2010). Figure 2 shows the difference in distributions over the STTS tags in the TüBa-D/Z (used for training) and in the entire TüBa-D/DC corpus. For example, one characteristic difference is: PPERS, the STTS tag for *irreflexive personal pronoun*, occurs with relative frequency of 2.14% in the TüBa-D/Z and with 6.29% in the TüBa-D/DC. There are two possible explanations: (i) the difference in relative frequency is due to tagging errors produced by the TreeTagger, or (ii) the difference in relative frequency is not due to tagging error, but rather an indication of the diverse nature of the two corpora. The subsequent evaluation will address this very question.

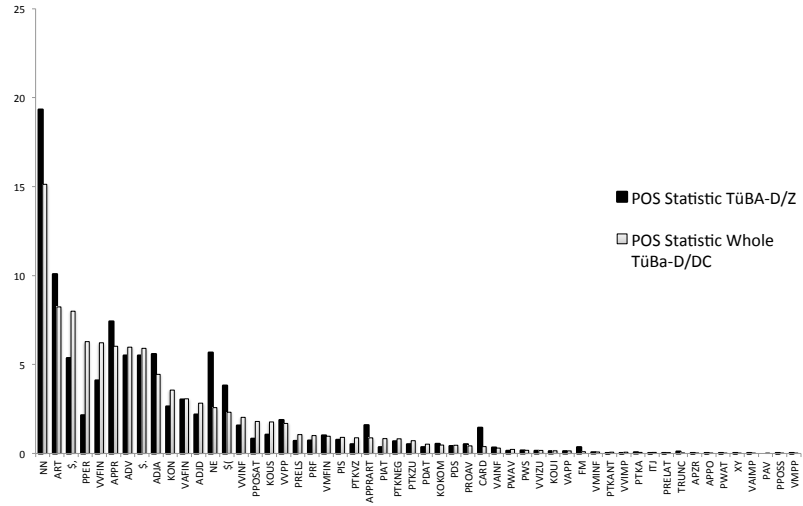

FIGURE 2 Distribution of part-of-speech tags in the TüBA-D/Z and the TüBa-D/DC

3.1 Evaluation Setup

Following Dipper (Dipper, 2010), we focus the evaluation of annotation quality on the part-of-speech layer and leave an evaluation of the parsing results to future work. Since the TüBa-D/DC is a very large resource, any manual evaluation can only be performed via data-sampling. The sampled data were chosen in such a way that they cover a long time span and represent materials of sufficient length. The selected data sample of six different subcorpora is summarized in Table

Author	**Titel**	**First Published**
Gottfried von Straßburg	Tristan	1210
Philipp Melanchthon	Die Augsburgische Konfession	1530
Abraham a Sancta Clara	Wunderlicher Traum von einem großen Narrennest	1703
Johann Wolfgang von Goethe	Die Leiden des jungen Werther	1774
Alexander von Humboldt	Kosmos	1845-1862
Theodor Däubler	Der Marmorbruch	1930

TABLE 3 Texts used for evaluation of the TüBa-D/DC

3.

The examples in (1), taken from the subcorpus Narrennest and (2), taken from the subcorpus Werther, give an impression of the different degrees of difficulty involved.

(1) *wann*(PWAV → KOUS) *solches*(PIAT → PIDAT) *Wunderwerck*(NN) *öffter*(ADV) *geschehe*(VVFIN) /($() *wie*(KOKOM → PWAV) *vil*(NE → ADV) *wurde*(VAFIN) *es*(PPER) *schwartze*(ADJA) *Larven*(NN) *absetzen*(VVINF) .($.)

'If such miracles happened more often / how much would it emit black larvae.'

(2) *um*(APPR) *ein*(ART) *Herz*(NN) *wie*(KOKOM) *das*(PDS → ART) *meine*(VVFIN → PPOSS) *zu*(PTKZU) *ängstigen*(VVINF) .($.)

'... in order to scare a heart like mine ...'

The part-of-speech tags in (1) and (2) were automatically assigned by the TreeTagger. Tagging errors occurred for those words which are followed by the pattern (<Wrong_Tag> → <Gold_Tag>). In the two examples, $(and $. are the STTS tags for sentence internal punctuation and sentence ending punctuation, respectively.

Example (1) contains four tagging mistakes. At least two of these are due to the diachronic nature of the text: in contemporary standard German, the word *wann* can only be used as an interrogative or relative pronoun (PWAV). Since the TreeTagger was trained on contemporary

Text	Accuracy	Av. Sent. Length
Tristan	68.9%	17.08
Die Augsburgische Konfession	88.6%	28.82
Wunderl. Traum von einem großen Narrennest	80.1%	37.09
Die Leiden des jungen Werther	98.7%	20.28
Kosmos	93.87%	32.38
Der Marmorbruch	97.45%	10.38

TABLE 4 Accuracy of the automatic part-of-speech tagging

newspaper texts, it assigns the PWAV tag instead of the KOUS tag (for: *subordinating conjunction*). This tagging mistake is typical for the subcorpus Narrennest where it occurs a total of 21 times and is due to the fact that *wann* used to have the meaning reserved in contemporary German for the conditional subordinating conjunction *wenn*. Due to its spelling, *vil*, which in contemporary German is spelled *viel*, seems to have been classified as an unknown word and is, thus, wrongly tagged as a proper name (NE). This in turn caused the mistagging of *wie* as a comparative partiticle (KOKOM) instead of the correct tag PWAV (for: *interrogative pronoun*). The most glaring tagging mistake in example (2) is the mistagging of *meine* as a final verb (VVFIN) instead of a possessive pronoun (PPOSS).

For each of the six data samples, the part-of-speech tags automatically assigned to the first ca. 13,000 tokens were manually inspected and corrected by an experienced research assistant. We will present both quantitative (subsection 3.2) and qualitative (subsection 3.3) results of the performed evaluation.

3.2 Evaluation Results

Table 4 shows the accuracy of the automatically assigned part-of-speech tags for every manually corrected subcorpus. The most striking result is that the accuracy for the Tristan subcorpus (68.9%) is substantially below the accuracy of all other subcorpora. However, this result is hardly surprising since Tristan is the only text that is written in Middle High German, a period usually dated from 1050 to 1350. The second lowest accuracy is obtained for Melanchthon's Augsburgische Konfession, a document written in Early New High German, a period dated from 1350 to 1650. The text authored by Abraham a Santa Clara belongs to the Barock period. In keeping with the artistic tradition of this period, it exhibits by far the highest average sentence length of all sample

subcorpora. The remaining Werther, Kosmos and Marmorbruch subcorpora belong to the New High German period and all have a higher tagging accuracy than the two subcorpora just discussed. The overall best accuracy of 98.7% of the Werther subcorpus is significantly higher than the tagging accuracy of 95.64% obtained on test data for the TüBa-D/Z, i.e. the corpus of newspaper text that the TreeTagger was trained on.

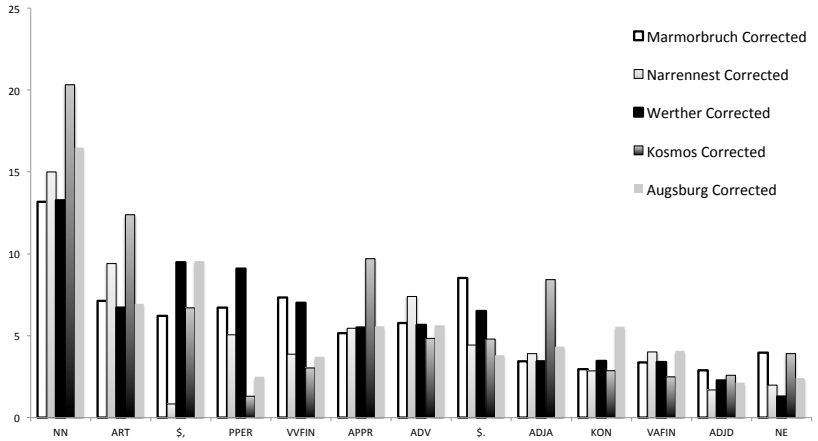

FIGURE 3 Relative Frequencies of hand-corrected (gold) tags

These findings raise a number of interesting questions: (i) The standard assumption that the performance of a part-of-speech tagger trained on out-of-domain material will always drop when applied to a new domain is only partially confirmed. (ii) For those subcorpora which surpass or are roughly equal to the tagging accuracy obtained for the TüBa-D/Z corpus, there seems to be some correlation with respect to sentence length. However, a closer inspection of the lexical and N-gram profiles of these texts would be beneficial in order to determine other relevant factors that may account for these resulting differences in tagging accuracy. As a first step in this direction, Figure 3 shows the relative frequency of the hand-corrected STTS tags for five of the six data samples.[9] The sixth data sample Tristan is excluded from the present evaluation because its error rate is so much higher than that of the other five data samples.

[9]For better readability, only the 13 STTS tags with the highest frequency in the whole TüBa-D/DC are shown in this and the following diagrams.

The following observations clearly stand out when looking at Figure 3: (i) While the relative frequency of each hand-corrected STTS tag differs substantially from sample to sample, there is typically one data sample where the frequency of such a hand-corrected tag is substantially higher than for the other subcorpora. For example, the tag APPR has relative frequency of 9.73% for the text Kosmos. (ii) The tag NN (for: *common noun*) has the highest relative frequency for all of them. (iii) Another discrepancy concerns the relative frequencies of punctuation tags such as the tag "$.". Here, the sample Marmorbruch exhibits the highest frequency which readily can be explained by the much shorter average sentence length of this subcorpus in comparison to the other texts (see Table 4 for the average sentence lengths). (iv) The two oldest texts differ from the rest by the substantially increased presence of the punctuation markers "$(". The four observations together give a first indication of the overall heterogeneity as well as of some sub-regularities of the different samples.

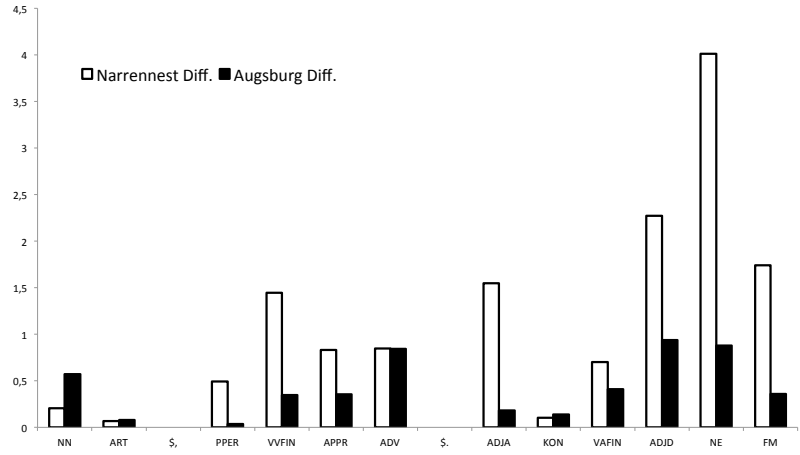

FIGURE 4 Automatically labeled vs. hand-corrected tags: differences in relative frequency

3.3 Error Analysis

Figure 4 provides an error profile for two of the six samples: Melanchton's Augsburg Konfession and Abraham a Santa Clara's Narrennest. These two texts were chosen because they have the highest number of tagging errors apart from the much older Tristan text. For each of the

STTS tags, Figure 4 shows the difference in relative frequency between the uncorrected and hand-corrected version of the data samples.[10] The figure reveals that this difference is highest for the tags NE for proper names, the two adjectival categories ADJA and ADJD, and the tags for the verbal categories VVFIN and VVPP. These results are in keeping with the received wisdom that nominal, verbal, and adjectival categories are hard to tag for German. More surprising is the discrepancy for foreign material which pertains particularly to the Narrennest data sample. The explanation is perhaps a surprising one and has to do with the orthography for prepositions such as *ohn*, which corresponds to modern German *ohne*, and *umb*, which corresponds to modern German *um*. Due to the non standard spelling, the guessing component of the tagger for unknown words apparently can not recognize them as German words and therefore treats them as foreign material.

Figure 4 shows that for most of the STTS tags included in the figure the difference between the automatically labeled and the hand corrected tag deviates significantly between the two subcorpora. This finding is at odds with the assumption that the two subcorpora share the same distribution of automatically assigned part-of-speech tags. Instead, the difference in behaviour of the two subcorpora suggests that they are, in fact, quite heterogenous in character.

A thorough analysis of all tagging errors would take the form of confusion matrices between all 55 STTS tags for each of the five data samples, included in the present evaluation and ultimately for the sixth data sample Tristan. This would go beyond the space limitations of the present paper. We will therefore focus on one STTS tag for the five data samples presently under consideration, namely the tag NN for common noun. Figure 5 shows which STTS tags are incorrectly used in place of the STTS tag NN and displays the relative frequencies for each of these erroneous classes of tags.

The highest amount of confusion arises with respect to the STTS tag NE. This is due to the fact that the unknown word guesser of the TreeTagger will typically treat an out-of-vocabulary NN as NE. The second most frequent confusion class concerns the erroneous tagging of nouns as adjectives. An inspection of the data reveals that differences in orthographic conventions (especially concerning capitalization) make up a common class of errors. In addition, there are also errors that are not due to the diachronic nature of the material, but would also occur in purely synchronic data sets. One such error is the mistagging of the

[10] In addition to the 13 STTS tags with the highest frequency in the TüBa-D/DC included in Figure 3, Figure 4 includes the STTS tag FM (for *foreign material*)

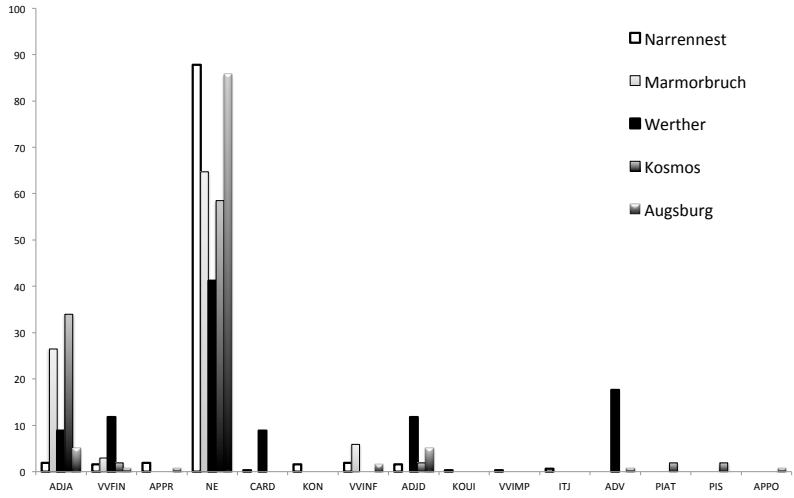

FIGURE 5 Wrongly annotated NNs

nominal head of a pre-head genitive phrase as an adjective (ADJA). An example of this kind is found in the Narrennest subcorpus: in *der Wüsten Heu-Schrecken*, the token *Wüsten* is mistagged as an ADJA. The erroneous tagging of nouns as VVFIN or VVINF makes up the third most frequent class of errors.

4 Accessing and Querying the TüBa-D/DC

The TüBa-D/DC data and its metadata are stored in a repository for digital data[11]. The metadata contained in this repository is publicly available to all and can be harvested automatically, but the corpus itself is restricted to the academic community. With the appropriate permissions, the user can download the TüBa-D/DC as a whole or individual texts from the repository.

There are several possibilities for querying the TüBa-D/DC and its metadata, ranging from web and desktop applications to programming language libraries. These methods vary in purpose, with some aimed at searching and retrieving the metadata and the corpus data, and others concentrating on programmatically retrieving and working with the data. Descriptions of the currently available software for accessing

[11] We are using the *Fedora Commons Repository* (Flexible Extensible Digital Object Repository Architecture), which is software to support long-term storage of large amounts of digital data and metadata, see www.fedora-commons.org

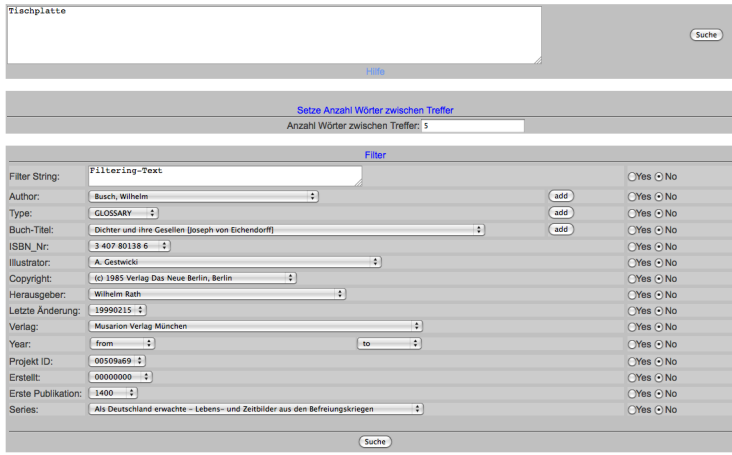

FIGURE 6 The user interface to the full text search engine, including metadata as filters

the TüBa-D/DC follow:

- **TüBa-D/DC-Search:** This search engine is a web application which allows both full-text search and metadata search. It makes use of Apache Lucene and Apache Solr (www.lucene.apache.org) for indexing and searching the corpus text and metadata. Several query formats are supported, including SRU/CQL[12] which is an open standard for distributed textcorpus applications. It allows the integration of the TüBa-D/DC-Search into other corpus query engines which support SRU/CQL. Filtering on metadata field values is also supported with the TüBa-D/DC-Search. For example, one can narrow the query to only texts of a particular author or a specified time range. It is then possible to view the resulting texts (including metadata) and the annotated texts (TCF). Work is currently underway to allow visualization and analysis of the annotated text.
- **CWB (Corpus Workbench):** *Corpus Workbench*[13] is a collection of tools developed at the IMS, University of Stuttgart, for managing and querying large annotated text corpora. The TüBa-D/DC has been converted to the required format, making it possible to take advantage of the whole suite of CWB tools.
- **TCFTools:** This is a Java library for parsing TCF data. It provides

[12]For more information on SRU/CQL, see http://www.loc.gov/standards/sru/cql-bibliographic-searching.html

[13]See http://cwb.sourceforge.net/ for more information about CWB.

a way of accessing the annotations within a Java program and offers full flexibility in accessing the linguistic information stored in a TCF file.

5 Conclusion and Future Work

We have described a diachronic corpus of German that uses selected materials from the German Gutenberg Project and enriches them with different linguistic annotation layers, including part-of-speech, lemmata, and constituent structure. Linguistic annotation is performed automatically by using statistical tools that have been trained with data from the Tübinger Baumbank des Deutschen (TüBa-D/Z). An evaluation of the POS tagging accuracy has revealed three common types of errors: (i) errors due to the diachronic nature of the corpus, such as differences in orthographic conventions, (ii) errors due to unknown words, and (iii) mistaggings (due to limitations of N-gram tagging) that would also occur in purely synchronic material. In order to improve POS performance, we plan to explore retraining of the POS tagger by using portions of the hand-corrected data samples as additional training material for the tagging model. This technique has already been applied successfully by Kübler and Baucom (2011). The second future task concerns an evaluation of the parsing accuracy. Here, we expect the same types of issues to arise that we have already encountered in the part-of-speech tagging evaluation.

Acknowledgments

We would like to thank three anonymous referees for their detailed and very helpful comments on an earlier version of this paper. Special thanks go to our colleagues Emanuel Dima, Ekaterina Kochmar and Julia Krivanek for their assistance in converting and annotating the GGP texts, Spyridoula Georgatou and Niko Schenk for enhancing the GGP metadata, as well as Silke Dutz and Heike Telljohann for creating the gold standard for the POS experiments. The research reported in this paper was funded by the D-SPIN and CLARIN-D projects, funded by the German Federal Ministry of Education and Research (BMBF).

References

Bikel, Daniel M. 2004. A distributional analysis of a lexicalized statistical parsing model. pages 182–189. Association for Computational Linguistics.

Dipper, Stefanie. 2010. Pos-tagging of historical language data: First experiments. pages 117–121.

Drach, Erich. 1937. *Grundgedanken der Deutschen Satzlehre*. Wissenschaftliche Buchgesellschaft.

Heid, Ulrich, Helmut Schmid, Kerstin Eckart, and Erhard Hinrichs. 2010. A corpus representation format for linguistic web services: The d-spin text corpus format and its relationship with iso standards. pages 494–499.

Herling, Simon Heinrich Adolf. 1821. Über die topik der deutschen Sprache. *Abhandlungen des frankfurterischen Gelehrtenvereins für deutsche Sprache* pages 296–362, 394.

Hinrichs, Marie, Thomas Zastrow, and Erhard Hinrichs. 2010. WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. pages 489–493.

Höhle, Tilman N. 1986. Der Begriff "Mittelfeld". Anmerkungen über die Theorie der topologischen Felder. *Kontroversen alte und neue. Akten des 7. Internationalen Germanistenkongresses Göttingen* pages 329–340.

Kübler, Sandra and Eric Baucom. 2011. Fast domain adaptation for part of speech tagging for dialogues. pages 41–48.

Schiller, Anne, Simone Teufel, and Christine Thielen. 1995. Guidelines für das Tagging deutscher Textcorpora mit STTS .

Telljohann, Heike, Erhard W. Hinrichs, and Sandra Kübler. 2004. The TüBa-D/Z treebank: Annotating german with a context-free backbone. pages 2229–2232.

Telljohann, Heike, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2009. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*.