Automatic Annotation and Manual Evaluation of the Diachronic German Corpus TüBa-D/DC

Erhard Hinrichs, Thomas Zastrow
University of Tübingen
Wilhelmstr. 19, 72074 Tübingen, Germany
[erhard.hinrichs, thomas.zastrow]@uni-tuebingen.de

Abstract

This paper presents the *Tübingen Baumbank des Deutschen Diachron* (TüBa-D/DC), a linguistically annotated corpus of selected diachronic materials from the German Gutenberg Project. It was automatically annotated by a suite of NLP tools integrated into WebLicht, the linguistic chaining tool used in CLARIN-D. The annotation quality has been evaluated manually for a subcorpus ranging from Middle High German to Modern High German. The integration of the TüBa-D/DC into the CLARIN-D infrastructure includes metadata provision and harvesting as well as sustainable data storage in the Tübingen CLARIN-D center. The paper further provides an overview of the possibilities of accessing the TüBa-D/DC data. Methods for full-text search of the metadata and object data and for annotation-based search of the object data are described in detail. The WebLicht Service Oriented Architecture is used as an integrated environment for annotation based search of the TüBa-D/DC. WebLicht thus not only serves as the annotation platform for the TüBa-D/DC, but also as a generic user interface for accessing and visualizing it.

Keywords: Diachronic text corpus, linguistic annotation, German

1. Motivation

This paper describes the Tübingen Baumbank des Deutschen Diachron (TüBa-D/DC), a linguistically annotated corpus of diachronic German. The TüBa-D/DC was created as part of the Common Language Resource and Technology Infrastructure initiative (CLARIN[1]) and is hosted by the CLARIN-D Center at the Eberhard Karls University of Tübingen.[2] CLARIN is an ESFRI infrastructure initiative that offers language resources and tools for scholars in the humanities and in the social sciences.[3]

The main motivation for the creation of the TüBa-D/DC was to create a sizable German text corpus that covers different stages of the language spanning from Middle High German to the present and that is linguistically annotated at different levels of analysis. To this end, materials from the German Gutenberg Archive[4] were selected and annotated by a suite of computational linguistics tools that are available as part of WebLicht[5], a service-oriented architecture for creating and executing customized chains of linguistic analysis (Hinrichs et.al. 2010). By including linguistic annotations in the TüBa-D/DC, it becomes possible to search the corpus materials in a much more fine-grained and targeted fashion than is possible in the case of an unannotated corpus. At the same time, the linguistic annotations can provide valuable information for the identification of subgenres included in the corpus and can also serve as cues for tracking language change.

The remainder of the paper is structured as follows: Section 2 describes the following aspects of the TüBa-D/DC: its textual basis (section 2.1), its linguistic annotations (section 2.2), its associated metadata (section 2.3), the assignment of persistent identifiers (section 2.4), the storage of the TüBa-D/DC in the CLARIN-D center repository, and means for metadata harvesting (section 2.5). Section 3 describes various methods and tools for accessing the TüBa-D/DC metadata and object data (section 3.1), as well as its linguistic annotations (section 3.2 to section 3.6). The paper concludes with a discussion of open research issues and of future directions for research (section 4).

2. The Tübinger Baumbank des Deutschen, Diachron (TüBa-D/DC)

2.1 Textual Basis

The TüBa-D/DC uses selected materials from the German Gutenberg Project (henceforth referred to as GGP) and enriches them with several linguistic annotation layers. Table 1 gives an overview of the size and the coverage of the TüBa-D/DC.

[1] CLARIN: www.clarin.eu

[2] CLARIN-D: http://clarin-d.net/index.php/en/

[3] CLARIN was recently granted the official status of a Research Infrastructure Consortium (ERIC). The CLARIN ERIC thereby enjoys the benefits of other international organizations such as tax exemption and administrative privileges.

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

[5] WebLicht: http://clarin-d.de/index.php/en/language-resources/weblicht-en

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: Dimensions of the TüBa-D/DC

Layer	**Tool used**
Tokens	OpenNLP Tokenizer
Part of speech (STTS)	TreeTagger
Lemmas	TreeTagger
Sentence boundaries	In-house tool
Named Entities (persons, locations, organizations and misc)	In-house tool
Constituent parse trees (TigerTB format)	Berkeley Parser

Table 2: Annotation layers of the TüBa-D/DC

Author	**Text**	**Accuracy**	**First Published**
Gottfried von Straßburg	Tristan	68.9%	1210
Philipp Melanchthon	Die Augsburgische Konfession	88.6%	1530
Abraham a Sancta Clara	Wunderl. Traum von einem großen Narrennest	80.1%	1703
Johann Wolfgang von Goethe	Die Leiden des jungen Werther	98.7%	1774
Alexander von Humboldt	Kosmos	93.87%	1845-1862
Theodor Däubler	Der Marmorbruch	97.45%	1930

Table 3: Data sample used for POS evaluation

2.2 Annotations

The different levels of annotation included in the TüBa-D/DC are listed in Table 2: tokenization, sentence boundary detection, part-of-specch tagging, lemmatization, named entity classification, and syntactic constituent structure.[6] For each annotation layer, Table 2 specifies the automatic tool that was used to produce the annotation. For all tools that require a trained data model, version 5.0 of the TüBa-D/Z (Tübinger Treebank des Deutschen/Zeitungstext)[7], a German treebank, was used as training data (Telljohann et al. 2004).

Figure 8 in section 3.5 shows an example of the constituent structure layer of the TüBa-D/DC. It displays the parse tree of the sentence *Die beste Bildung findet ein gescheiter Mensch auf Reisen* ('The best education an intelligent person receives during travels') taken from Goethe's novel *Wilhelm Meister's Lehrjahre*. This novel is included in the GGP, and its syntactic structure is generated by the Berkeley parser. 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). The phrasal nodes include a layer of topological fields such as VF (*Vorfeld*), LK (*Linke Klammer*), and MF (*Mittelfeld*). Topological fields in the sense of Höhle (1986) 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. Detailed information about the syntactic annotation scheme used in the TüBa-D/DC and in the TüBa-D/Z can be found in Telljohann et al. (2009).

Since the linguistic annotation of the TüBa-D/DC was performed automatically, the resulting annotations are not 100% accurate. In order to assess the quality of the annotation, an evaluation of the POS tagging accuracy was performed via data-sampling. The selected data sample of six different subcorpora (see Table 3**)** was chosen in such a way that they cover a long time span.

For each of the six data samples, the part-of-speech tags automatically assigned to the first 13,000 tokens were manually inspected and corrected by an experienced research assistant. The error analysis 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.[8]

2.3 Metadata

All texts from the GGP come with a set of metadata in a Dublin Core-like format. This metadata contains information about the individual texts such as author, year of publication etc.

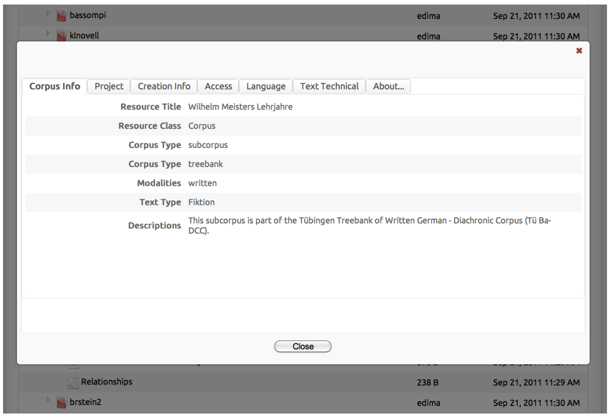
In the GGP, the set of metadata elements is not uniform across the data entries. For example, some of the metadata sets contain an entry for the year of first publication, while others don't. This is partly due to the fact that it is not always easy, and in some cases not even possible, to determine the correct data. Another possible contributing factor to the incompleteness of the metadata may be that the GGP is an ongoing project that enlists the help of thousands of volunteers, which

[6] Dependency parsing is in preparation.

[7] http://www.sfs.uni-tuebingen.de/tuebadz.shtml

[8] See Hinrichs and Zastrow (2012) for a more detailed discussion.

can degrade uniformity[9]. It was therefore necessary to manually check and correct, where possible, the metadata of all 19,377 texts in the TüBa-D/DC. After the manual correction, the metadata was then converted from its original format to the *Component MetaData Infrastructure* (CMDI)[10] format, the metadata standard used in CLARIN. Figure 1 shows an excerpt of CMDI metadata, describing Goethe's novel *Wilhelm Meisters Lehrjahre*.



bassompi	edima	Sep 21, 2011 11:30 AM
klnovell	edima	Sep 21, 2011 11:30 AM

Resource Title	Wilhelm Meisters Lehrjahre
Resource Class	Corpus
Corpus Type	subcorpus
Corpus Type	treebank
Modalities	written
Text Type	Fiktion
Descriptions	This subcorpus is part of the Tübingen Treebank of Written German - Diachronic Corpus (Tü Ba-DCC).

Relationships	238 B	Sep 21, 2011 11:29 AM
brstein2	edima	Sep 21, 2011 11:30 AM

Figure 1: Metadata Goethe

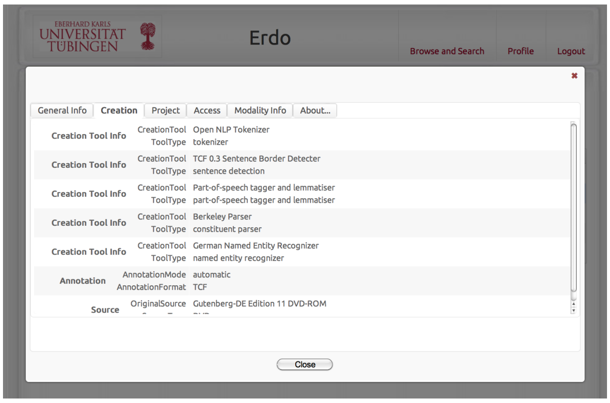
In addition to the CMDI files for the individual subcorpora, there are two additional CMDI files. One is the master document which contains advanced and detailed information about the TüBa-D/DC as a whole. This includes information about the linguistic layers as well as the process of creation the TüBa-D/DC. This master document provides the first point of access by one who is interested in information about the TüBa-D/DC. An excerpt of this master document is shown in Figure 2. Another metadata document, which is linked from the master document described above, contains links to all of the 19,377 subcorpora of the TüBa-D/DC.

2.4 Persistent Identifiers

Metadata can be harvested from distributed catalog systems. This harvested metadata is linked to the actual object data via *Persistent Identifiers* (PID). A PID guarantees that a resource which is marked with it can be found even when its URL changes. A *resolver* maps the PID to one or more external URLs[11]. Several PID and resolving services are available. The Tübingen CLARIN-D center makes use of the *Handle System*[12], which was implemented by the EPIC consortium[13] and is operated by the GWDG[14] within Germany.

In the case of the TüBa-D/DC, its PID points to the master CMDI document, discussed above, which contains links to all the 19,377 individual documents of the corpus. Although it would also be possible to assign a PID to each individual text, or to go even further, to every sentence or even every individual token of the corpus, this was not considered practical in the case of the TüBa-D/DC. But in principle it is possible to define an arbitrary depth of granularity for every resource individually.

2.5 Storage and Availability



Creation Tool Info	CreationTool	Open NLP Tokenizer
	ToolType	tokenizer
Creation Tool Info	CreationTool	TCF 0.3 Sentence Border Detecter
	ToolType	sentence detection
Creation Tool Info	CreationTool	Part-of-speech tagger and lemmatiser
	ToolType	part-of-speech tagger and lemmatiser
Creation Tool Info	CreationTool	Berkeley Parser
	ToolType	constituent parser
Creation Tool Info	CreationTool	German Named Entity Recognizer
	ToolType	named entity recognizer
Annotation	AnnotationMode	automatic
	AnnotationFormat	TCF
Source	OriginalSource	Gutenberg-DE Edition 11 DVD-ROM

Figure 2: Metadata master document

In order to assure its sustainability, the TüBa-D/DC is stored in the data repository of the Tübingen CLARIN-D center. This repository uses Fedora Commons, a general purpose open-source repository system, as a backend for storing digital objects. In order to facilitate data management, the Tübingen repository makes use of *Extensible Repository System of Digital Objects* (ERDO), a graphical user interface developed in Tübingen for depositing and displaying digital objects.[15] Figures 1 and 2 exemplify the ERDO functionality for visualization of CMDI metadata.

While the metadata is freely available and can be harvested as described above, the object data stored in digital objects is accessible from within the CLARIN identity federation.

[9] Goethe's novel *Die Leiden des jungen Werther*, one of the texts contained in the TüBa-D/DC, 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. 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.

[10] More information on the CMDI metadata format can be found here: http://www.clarin.eu/cmdi

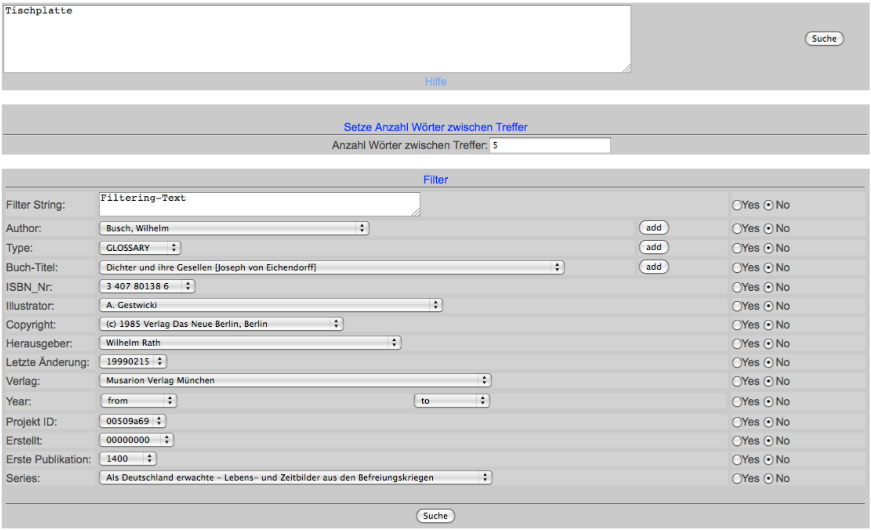
[11] The PID for the TüBa-D/DC is http://hdl.handle.net/11858/00-1778-0000-0001-DDAF-D

[12] http://www.handle.net/

[13] http://www.pidconsortium.eu/

[14] http://handle.gwdg.de:8080/pidservice/

[15] ERDO was developed by the Tübingen CLARIN-D center. See Dima et al. (2012) for more information.



Tischplatte

Suche

Hilfe

Setze Anzahl Wörter zwischen Treffer

Anzahl Wörter zwischen Treffer: 5

Filter

Filter String:	Filtering-Text		○ Yes ● No
Author:	Busch, Wilhelm	add	○ Yes ● No
Type:	GLOSSARY	add	○ Yes ● No
Buch-Titel:	Dichter und ihre Gesellen [Joseph von Eichendorff]	add	○ Yes ● No
ISBN_Nr:	3 407 80138 6		○ Yes ● No
Illustrator:	A. Gestwicki		○ Yes ● No
Copyright:	(c) 1985 Verlag Das Neue Berlin, Berlin		○ Yes ● No
Herausgeber:	Wilhelm Rath		○ Yes ● No
Letzte Änderung:	19990215		○ Yes ● No
Verlag:	Musarion Verlag München		○ Yes ● No
Year:	from to		○ Yes ● No
Projekt ID:	00509a69		○ Yes ● No
Erstellt:	00000000		○ Yes ● No
Erste Publikation:	1400		○ Yes ● No
Series:	Als Deutschland erwachte – Lebens– und Zeitbilder aus den Befreiungskriegen		○ Yes ● No

Suche

Figure 3: The Lucene search interface

3. Accessing the TüBa-D/DC

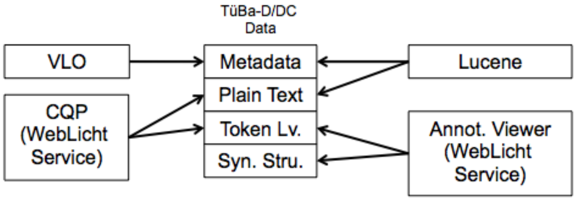


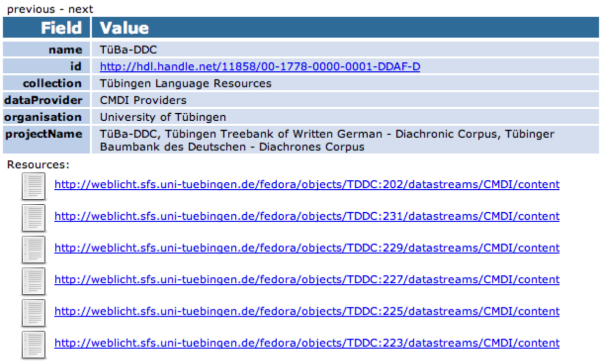
Figure 4: Accessing the TüBa-D/DC

Figure 4 shows an overview of the possibilities of accessing the TüBa-D/DC. The individual methods for full-text search of the metadata and object data and for annotation-based search of the object data are described in detail below.

3.1 Metadata & Full Text Search

The use of the CMDI metadata format makes it possible to harvest and integrate the TüBa-D/DC metadata into catalogue systems such as the *Virtual Language Observatory* (VLO)[16], maintained by CLARIN and hosted by the Max Planck Institute for Psycholinguistics in Nijmegen. The VLO is a web-based application with facetted search functionality for browsing and displaying metadata for language resources and tools. The VLO also provides functionality for full-text search and for displaying the site at which a resource or tool resides (see Figure 5).

Another way of accessing the metadata and the full text of the TüBa-D/DC is via the Lucene full text search engine, hosted at the CLARIN-D center in Tübingen. This application allows a user to search, for example, for specific words using Lucene's query language. In addition, it is possible to filter the results with the help of the metadata, for example to limit the results to a specific time frame (see Figure 3).



previous - next

Field	Value
name	TüBa-DDC
id	http://hdl.handle.net/11858/00-1778-0000-0001-DDAF-D
collection	Tübingen Language Resources
dataProvider	CMDI Providers
organisation	University of Tübingen
projectName	TüBa-DDC, Tübingen Treebank of Written German - Diachronic Corpus, Tübinger Baumbank des Deutschen - Diachrones Corpus

Resources:

- http://weblicht.sfs.uni-tuebingen.de/fedora/objects/TDDC:202/datastreams/CMDI/content
- http://weblicht.sfs.uni-tuebingen.de/fedora/objects/TDDC:231/datastreams/CMDI/content
- http://weblicht.sfs.uni-tuebingen.de/fedora/objects/TDDC:229/datastreams/CMDI/content
- http://weblicht.sfs.uni-tuebingen.de/fedora/objects/TDDC:227/datastreams/CMDI/content
- http://weblicht.sfs.uni-tuebingen.de/fedora/objects/TDDC:225/datastreams/CMDI/content
- http://weblicht.sfs.uni-tuebingen.de/fedora/objects/TDDC:223/datastreams/CMDI/content

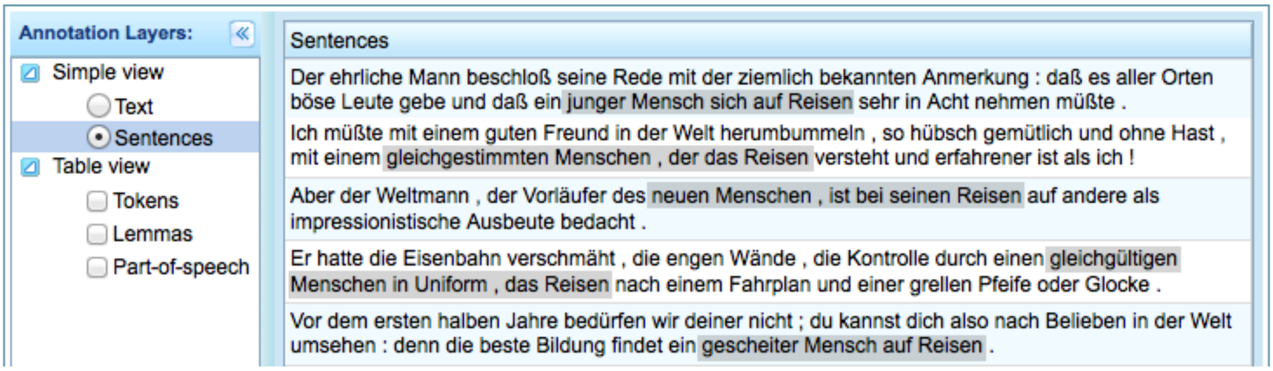
Figure 5: A screenshot of the VLO, displaying a CMDI metadata record about the TüBa-D/DC and its subcorpora

3.2 Annotation-Based Search

The linguistic annotations in the TüBa-D/DC allow for much more sophisticated queries of the data than is possible via plain text and metadata search. For example, the lemmatization layer supports queries not only to individual word forms, but to all occurrences of a particular lemma in the corpus. The part of speech annotation layer can be searched for recurring N-Grams, and the syntactic constituent structure for particular phrase types or syntatctic constructions.

The WebLicht Service Oriented Architecture is used as an integrated environment for annotation based search of the TüBa-D/DC. WebLicht thus not only serves as the annotation platform for the TüBa-D/DC, but also as the user interface for the CQP query engine.

[16] The VLO can be found here: http://catalog.clarin.eu/ds/



Annotation Layers:

☑ Simple view

○ Text

● Sentences

☑ Table view

□ Tokens

□ Lemmas

□ Part-of-speech

Sentences

Der ehrliche Mann beschloß seine Rede mit der ziemlich bekannten Anmerkung : daß es aller Orten böse Leute gebe und daß ein junger Mensch sich auf Reisen sehr in Acht nehmen müßte .

Ich müßte mit einem guten Freund in der Welt herumbummeln , so hübsch gemütlich und ohne Hast , mit einem gleichgestimmten Menschen , der das Reisen versteht und erfahrener ist als ich !

Aber der Weltmann , der Vorläufer des neuen Menschen , ist bei seinen Reisen auf andere als impressionistische Ausbeute bedacht .

Er hatte die Eisenbahn verschmäht , die engen Wände , die Kontrolle durch einen gleichgültigen Menschen in Uniform , das Reisen nach einem Fahrplan und einer grellen Pfeife oder Glocke .

Vor dem ersten halben Jahre bedürfen wir deiner nicht ; du kannst dich also nach Belieben in der Welt umsehen : denn die beste Bildung findet ein gescheiter Mensch auf Reisen .

Figure 6: Visualization of CQP query result in WebLicht

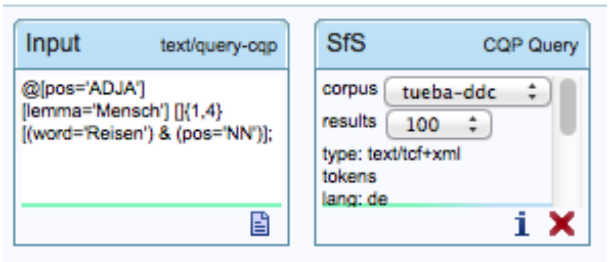
3.3 Lemma & POS based Search

The *Corpus Query Processor* (CQP) is part of the *Corpus Workbench* (see Christ et al. 1994 for more details). CQP is widely used in computational linguistics to index and query large text corpora. With its integrated query language, linguistic annotations on the token level (the token itself, lemma or part of speech annotations) can be queried. With CQP, it is unfortunately not possible to query the syntax-level annotations.

The following CQP query demonstrates how to use token, part of speech and lemma information in a search. It searches for constructs in which any token with the part of speech tag *ADJD* is followed by a token with the lemma *Mensch*. Then, a gap of one to four tokens is allowed after which a token *Reisen* with the part of speech tag *NN* should be present:

```
@[pos='ADJA'] [lemma='Mensch']
[]{1,4} [(word='Reisen') &
(pos='NN')];
```

Depending on the configured context, CQP returns the results as „Keyword in Context“ (KWIC) or the entire sentence(s) in which a match was found. For example, in the TüBa-D/DC, the query above will return 9 results (see Figure 6).



Input text/query-cqp

@[pos='ADJA'] [lemma='Mensch'] []{1,4} [(word='Reisen') & (pos='NN')];

SfS CQP Query

corpus: tueba-ddc

results: 100

type: text/tcf+xml

tokens

lang: de

Figure 7: Querying the TüBa-D/DC from within WebLicht

3.4 CQP Integration into WebLicht

CQP uses a command-line interface, which makes it simple to integrate it via a web service wrapper into WebLicht. The concrete integration into WebLicht is shown in Figure 7. Using WebLicht's chaining capabilities, the user can first write a query in CQP syntax. Parameters for the service can be selected in the service box. The parameters for this service include a choice of corpus and the maximum number of sentences to return. It should be mentioned that in this particular implementation, the CQP engine will always give back full sentences. The results are returned in WebLicht's processing format *Text Corpus Format* (TCF, see Heid et al. 2010)[17], which is an easy to use, standard-compliant XML format. Using this format has the advantage that additional WebLicht tools can be subsequently applied to the CQP output, including additional linguistic annotations, statistical analysis, and visualizations, as well as converters to other linguistic data formats.

3.5 Visualization of Syntactic Structures

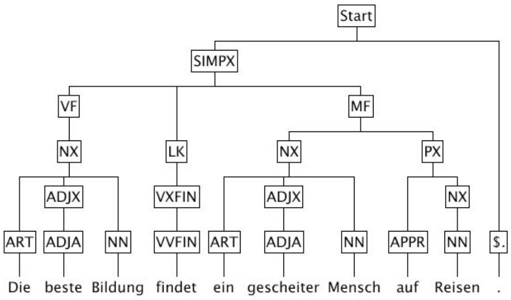


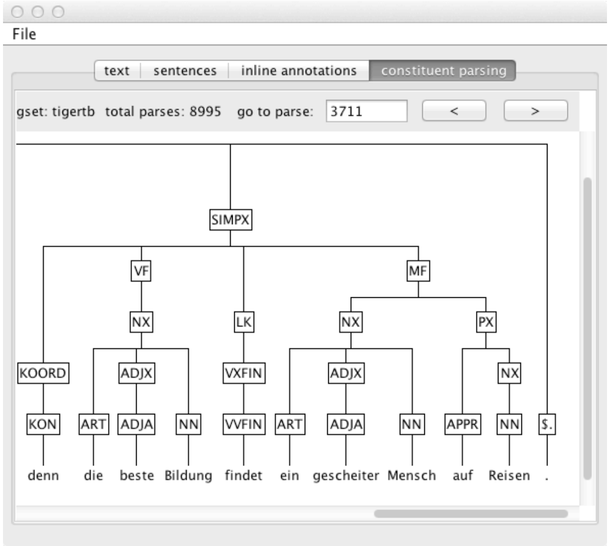
Figure 8: Visualization of a parse tree with WebLicht's Annotation Viewer

In addition to the token level, the TüBa-D/DC contains syntactic structures in the form of constituent parse trees. These parse trees cannot be accessed via CQP, but WebLicht can fill the gap here with its integrated

[17] Detailed information about the TCF format can be found online: http://clarin-d.de/index.php/en/language-resources/weblicht-en/tutorials

Annotation Visualizer. With this visualization software one can take a look at these syntactic structures as graphical parse trees, as are commonly used in computational linguistics (see Figure 8). The token-based information can also be viewed in a table-like format.

3.6 Tviewer for Non-Web-Based Access



gset: tigertb total parses: 8995 go to parse: 3711

Figure 9: TViewer for visualization

The TüBa-D/DC consists of 19,377 individual subcorpora. Most of these subcorpora are too large to be displayed as a whole from within a web application like WebLicht. Here, the Tviewer[18] can be used: it is a desktop version of the web-based Annotation Visualizer described above. The application is written in Java and can be executed on any modern operating system (see Figure 9).

4. Conclusion and Further Work

The TüBa-D/DC is already integrated into WebLicht. In a next step, it will be also available via CLARIN's Federated Search. Adding more linguistic annotations to the TüBa-D/DC, for example dependency parse trees and coocurrences, is also planned.

5. Acknowledgement

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 carried out within the D-SPIN and CLARIN-D projects, funded by the German Federal Ministry of Education and Research (BMBF).

6. References

Christ, Oli. 1994. A modular and flexible architecture for an integrated corpus query system. *COMPLEX'94*, Budapest.

Dima, Emanuel, Verena Henrich, Erhard Hinrichs, Marie Hinrichs, Christina Hoppermann, Thorsten Trippel, Thomas Zastrow and Claus Zinn. 2012. A Repository for the Sustainable Management of Research Data. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12).

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. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC'10), pages 494–499.

Hinrichs, Erhard and Thomas Zastrow. 2012. Linguistic Annotations for a Diachronic Corpus of German. *Linguistic Issues in Language Technology*, Vol. 7.7.

Hinrichs, Marie, Thomas Zastrow, and Erhard Hinrichs. 2010. WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC'10), 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 29–340.

Schiller, Anne, Simone Teufel, and Christine Thielen. 1995. *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Technical Report. Institut fur maschinelle Sprachverarbeitung, Stuttgart.

Telljohann, Heike, Erhard W. Hinrichs, and Sandra 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), 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). Technical Report. Seminar für Sprachwissenschaft, University of Tübingen.

[18] http://clarin-d.de/index.php/en/language-resources/weblicht-en/tutorials