TüBa-D/Z Release 10.0

The TüBa-D/Z treebank is a manually syntactically annotated German newspaper corpus based on data taken from the daily issues of 'die tageszeitung' (taz). The treebank currently comprises 3,644 newspaper articles (95,595 sentences; 1,787,801 tokens).

The TüBa-D/Z treebank is available in several formats. Not all formats support all types of annotation. The following table summarizes the annotations included in each format:

	Negra Export[1]			CoNLL			Penn		ExportXML	Chunks
	export3	export4	TigerSearch export4	2006	2010	2011/2012	v1	v2	v2	
POS	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Syntax	✓	✓	✓			✓	✓	✓[2]	✓	
Markup of Typos	✓	✓	✓						✓	
Morphology	✓	✓	✓	✓	✓				✓	
Anaphora / Coreferences	✓	✓	✓		✓	✓			✓	
Lemmas	✓	✓	✓	✓	✓	✓			✓	
Named Entities	✓	✓	✓		✓	✓	✓	✓	✓	✓
Discourse Relations	✓	✓	✓						✓	
Word Sense	✓	✓	✓			✓			✓	
Dependency Syntax[3]				✓	✓				✓	
Chunks[3]										✓
File Name	tuebadz-10.0-export3.txt.zip	tuebadz-10.0-export4.txt.zip	tuebadz-10.0-export4-tiger.txt.zip	tuebadz-10.0-conll2006.txt.zip	tuebadz-10.0-conll2010.txt.zip	tuebadz-10.0-conll2011.txt.zip	tuebadz-10.0-penn-v1.txt.zip	tuebadz-10.0-penn-v2.txt.zip	tuebadz-10.0-exportXML-v2.xml.zip	tuebadz-10.0-chunks.txt.zip
Encoding	ISO-8859-1	ISO-8859-1	ISO-8859-1	UTF-8	UTF-8	UTF-8	UTF-8	UTF-8	UTF-8	UTF-8

1. The export3 and export4 formats differ only in the location of the lemma information. In export3 lemma information is contained in the comment field. In export4 lemmas have their own column. Several long sentences are removed in the TigerSearch export4 format which would otherwise prevent the treebank from being loaded into TigerSearch.
2. No unattached phrases
3. Automatically generated

Files included in this release:

tuebadz-10.0-export3.txt	This is the Negra Export 3 format, with lemmas in the comment field.
tuebadz-10.0-export4.txt	This is the Negra Export 4 format, with lemmas in their own column.
tuebadz-10.0-export4-tiger.txt	This format can be used with TigerSearch. Approximately 50 sentences are removed from this data since they are too long for TigerSearch to handle.
tuebadz-10.0-conll2006.txt	This format was used in a CoNLL shared task in 2006. See below for details about column usage for the TüBa-D/Z.
tuebadz-10.0-conll2010.txt	This format was used in a CoNLL shared task in 2010. See below for details about column usage for the TüBa-D/Z.
tuebadz-10.0-conll2011.txt	This format was used in CoNLL shared tasks in 2011 and 2012. See below for details about column usage for the TüBa-D/Z.
tuebadz-10.0-penn-v1.txt	Penn Treebank format.
tuebadz-10.0-penn-v2.txt	Penn Treebank format with no unattached phrases.
tuebads-exportXML-v2.xml	This is an XML representation containing all of the information also contained in the export format. See here for more information about this format, and access to a java API for reading/writing. Please note that the java API was written and is maintained externally.
tuebadz-10.0-chunks.txt	A chunked version of the treebank. See here for more information about the chunking annotation.
tuebadz-10.0-README.pdf	This readme.

The negra export format can be used in combination with the annotation tool Annotate (no longer maintained), which was developed in the Project negra at the Computational Linguistics Department at the University of the Saarland or with the TIGERSearch Tool developed in the TIGER project at the Institute for Natural Language Processing, University of Stuttgart.

TüBaD/Z CoNLL 2006/2010/2011 Formats

The 2006/2010/2011 CoNLL format outputs are automatically generated from the EXML format of TübaD/Z, which follow the corresponding CoNLL formats used in the original CoNLL shared tasks. The original descriptions of each column can be found at:

- ConLL 2006: http://ilk.uvt.nl/conll/#dataformat
- CoNLL 2010: http://stel.ub.edu/semeval2010-coref/datasets
- CoNLL 2011 (same as 2012): http://conll.cemantix.org/2012/data.html

Certain information in some columns are either absent or not applicable to the TüBa-D/Z corpus, which leads to differences in some of the column definitions. The following aims to document the specifics.

CoNLL 2006

Each row that represents a word token has exactly 10 columns.

Column 9-10: always marked as "_", since they are absent in the TüBa-D/Z.

CoNLL 2010

Each row that represents a word token has exactly 17 columns.

To be consistent with the format definition, column 4, 6, 8, 10, 12 and 14 are "predicted" values, thus marked as "_".

Column 15-16: always marked as "_", since they are absent in the TüBa-D/Z.

There are no "N+M" columns after column 16. These are meant for semantic role labeling that were to be derived from 15 and 16 in the shared task. Since 15 and 16 themselves are filled with "_", due to their absence in Tüba D/Z, these columns do not exist in this data.

Column 17 is always Coreference.

CoNLL 2011/2012

Each row that represents a word token has exactly 13 columns. Some of the columns vary somewhat from the official definition:

Column 1 (Document ID): the newspaper article id in the form TYYMMDD.articleNumber

Column 2 (Part Number): the GLOBAL sentence ID. Numbering does not restart within each document, therefore the part number corresponds to the sentence ID in the treebank (ranging from 1 to 95595 for Release 10.0). Thus a document should be solely identified by the doc ID.

Column 7 (Predicate lemma): the lemma of every token is represented

Column 8 (Predicate Frameset ID): filled with "-", as the TüBa-D/Z does not have this information

Column 9 (Word sense): we use GermanNet IDs to represent word senses

Column 10 (Speaker): filled with "-", as the TüBa-D/Z does not have this information

Column 12 (Predicate Arguments): is filled with "-", as columns 12:N are derived from the predicate lemma in column 7, which is not available.

Column 13 (Coreference): coreference information is always in column 13