TüBa-D/DP Stylebook

Release 5

Daniël de Kok and Sebastian Pütz

1 Introduction

TüBa-D/DP is a machine-annotated dependency treebank of German. The goal of TüBa-D/DP is to offer high-qualitity syntactic annotations for a huge amount of contemporary German text. The annotations follow the TüBa-D/Z UD annotation guidelines (Çöltekin et al. 2017) as closely as possible. TüBa-D/DP currently consists of the subcorpora summarized in Table 1.

Table 1: Subcorpora of the TüBa-D/DP.

Subcorpus	**Genre**	**Sentences**	**Tokens**
Europarl (Koehn 2005; Tiedemann 2012)	Parliamentary proceedings	2.2M	55M
taz (1986-2009)	Newspaper	23.2M	397.3M
Wikipedia (2020)	Encyclopedia	45.5M	917.5M
Political speeches (Barbaresi 2018)	Speeches held by German officials	619,152	12.8M

TüBa-D/DP is provided in the CoNLL-U format[1] and provides the following annotations layers listed in Table 2.

Table 2: Annotation layers of the TüBa-D/DP.

Layer	**CoNLL-U column**	**Attribute**
Universal POS	UPOS	
STTS POS	XPOS	
Lemma	LEMMA	
UD morphology	FEATS	
Dependency head	HEAD	
Dependency relation	DEPREL	

[1] https://universaldependencies.org/format.html

Layer	CoNLL-U column	Attribute
Topological field	MISC	TopoField
TüBa-D/Z morphology	MISC	Morph
Named entity	MISC	NE

The differences between the TüBa-D/DP and TüBa-D/Z UD annotation schemes are described in Section 2. The annotation tools that ere used are described in Section 3.

2 Deviations from TüBa-D/Z UD annotations

2.1 Preposition-determiner contractions

Contractions of determiner and a preposition, such as *zur* (*zu der*) and *am* (*an dem*) are split into two tokens in the TüBa-D/Z UD. For instance, the token *am* in the sentence *Der Mann wurde noch am Tatort festgenommen.* is encoded as follows:[2]

```
5-6     am
5       an
6       dem
```

The preposition and determiner are represented as separate tokens `5` and `6`. The token span `5-6` represents the original token. In the TüBa-D/DP, we do not split such contractions to simplify processing. Thus, the token would be represented as follows in the TüBa-D/DP:

```
5       am
```

2.2 Lemmas

2.2.1 Determiners

Due to the ambiguity in lemmatization of articles and relative pronouns, articles and relative pronouns are lemmatized as respectively *d* and *e* for definite and indefinite forms. For example:

- *den* → *d*
- *einem* → *e*
- *dessen* → *d*

2.2.2 Personal and possesive pronouns

Personal and possesive pronouns are lemmatized as in Table 3.

[2] The token annotations are removed for brevity.

Table 3: Lemmatization of personal and possesive pronouns.

Lowercased forms	**Lemma**
ich, mich, mir, meiner	*ich*
du, dir, dich, deiner	*du*
er, ihn, ihm, seiner	*er*
sie, ihr, ihnen, ihrer	*sie*
es, 's	*es*
wir, uns, unser	*wir*
ihr, euch	*ihr*

2.2.3 Indefinite pronouns

Indefinite pronouns (PIAT, PIDAT, PIS) show ambiguities in form-lemma mappings. For these categories, forms are truncated to a common prefix. Table 4 lists example tranformations with forms taken from TüBa-D/Z.

Table 4: Lemmatization of indefinite pronouns.

Lowercased forms	**Lemma**
jeder, jede, jedes, jede(r), jeden, jede/r, jedem	*jed*
solche, solchen, solcher	*solch*
einige, einiges, einiger, einigen	*einig*
jedwedem, jedweden, jedwedes, jedweder	*jedwed*
vieler, vielen, viel, viele, vielem	*viel*
meisten, meiste	*meist*

2.2.4 Separable verb prefixes

TüBa-D/DP marks separable verb prefixes as in TüBa-D/Z. For example, the inflected form *abgezeichnet* is lemmatized as *ab#zeichnen*. This type of transformation prefers analyses with longer prefixes over shorter prefixes. For instance, *hinzugefügt* is lemmatized as *hinzu#gefügt*, and not as *hin#zu#gefügt*.

Separated prefixes are also taken into account. For example, *zeichnen* in

> Diese änderungen zeichnen sich bereits ab .

is also lemmatized as *ab#zeichnen*.

In some cases, conjunctions of separable prefixes are also annotated. For example, *nimmt* in

> [...] nimmt eher zu als ab

is lemmatized as *zu#nehmen|ab#nehmen*. However, the post-processing rules for such conjunctive cases may not be exhaustive.

2.3 Topological fields

Since dependency grammar does not use phrasal nodes, topological fields are annotated on a token-level (Kok and Hinrichs 2016). Each token has an feature *TopoField* that marks the field that the token is in.

Topological fields are annotated in TüBa-D/Z UD version on tokens as a list of all the topological fields that the token participates in. For instance, the topological field annotation *TopoField=NF-VF-MF* indicates that a token is in the *middle field* (MF) of the most specific clause and in the *final field* (NF) of the most general clause. In the TüBa-D/DP, we only annotate the most specific field, thus in this case *MF*.

2.4 Named entities

For improving quality in sequence labeling, the named entities are converted to an *IOB*-scheme. Where *B-* marks the first word in a named entity, *I-* subsequent words in a named entity, and *O* words that are not part of a named entity.

In the TüBa-D/Z UD corpus, named entities can be hierarchical. For instance, *FC St. Pauli* is annotated as *ORG* (organization); furthermore *St. Pauli* is annotated as *LOC* (location). In such cases, we take the outer annotation, so these tokens would be annotated as `FC/B-ORG St./I-ORG Pauli/I-ORG`.

3 Annotation tools

TüBa-D/DP was annotated with the following tools:

- **Tokenization**:
  - Wikipedia: SoMaJo (Proisl and Uhrig 2016)
  - taz: TüPP-D/Z tokenizer (Ule 2004)
- **Annotation layers**: sticker2 (Kok, Falk, and Pütz 2020)

References

Barbaresi, Adrien. 2018. "A Corpus of German Political Speeches from the 21st Century." In.

Çöltekin, Çağrı, Ben Campbell, Erhard Hinrichs, and Heike Telljohann. 2017. "Converting the TüBa-d/Z Treebank of German to Universal Dependencies." In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, 27–37. Gothenburg, Sweden: Association for Computational Linguistics. https://www.aclweb.org/anthology/W17-0404.

Koehn, Philipp. 2005. "Europarl: A Parallel Corpus for Statistical Machine Translation." In *Conference Proceedings: the tenth Machine Translation Summit*, 79–86. Phuket, Thailand: AAMT; AAMT.

Kok, Daniël de, Neele Falk, and Tobias Pütz. 2020. "Sticker2: A Neural Syntax Annotator for Dutch and German."

Kok, Daniël de, and Erhard Hinrichs. 2016. "Transition-Based Dependency Parsing with Topological Fields." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2:1–7.

Proisl, Thomas, and Peter Uhrig. 2016. "SoMaJo: State-of-the-Art Tokenization for German Web and Social Media Texts." In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, 57–62. Berlin: Association for Computational Linguistics (ACL). http://aclweb.org/anthology/W16-2607.

Tiedemann, Jörg. 2012. "Parallel Data, Tools and Interfaces in OPUS." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2214–8. Istanbul, Turkey: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Ule, Tylman. 2004. "Markup Manual for the Tübingen Partially Parsed Corpus of Written German (TüPP-d/Z)." In *Sonderforschungsbereich 441, Seminar Für Sprachwissenschaft, Universität Tübingen*, 28:2006.