TüBa-D/DP Stylebook

Release 4

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. TüBa-D/DP attempts to provide familiar annotations by following the TüBa-D/Z annotation guidelines (Telljohann et al. 2006) 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	Parliamentary proceedings	2.2M	55M
taz (1986-2009)	Newspaper	29.9M	393.7M
Wikipedia (2019)	Encyclopedia	42.2M	849.5M
Common Crawl (2019)	Webpages	1.4B	27.3B

TüBa-D/DP is provided in the CoNLL-X format (Buchholz and Marsi 2006) and provides the following annotations layers: part-of-speech tags, morphology, lemmas, topological fields, and dependency relations.

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

2 Differences to TüBa-D/Z annotations

2.1 Morphology

The TüBa-D/Z morphological tags were expanded to attribute-value pairs. For example the TüBa-D/Z morphological tag `3sit` is encoded as `mood:indicative|number:singular|person:3|tense:past` in TüBa-D/DP. The morphological tag expansions make quering TüBa-D/DP on specific morphological attributes simpler and improves automatic morphological annotation.

The morphological attributes and their possible values are shown in Table 2. Table 3 shows the morphological attributes for each part-of-speech.

Table 2: Morphological attributes and values.

Attribute	**Value**
case	nominative
	genetive
	dative
	accusative
gender	masculine
	feminine
	neuter
number	plural
	singular
mood	indicative
	subjunctive
person	1
	2
	3
tense	present
	past

Table 3: Morphological attributes of each part-of-speech tag.

Part-of-speech	**Attributes**
ADJA	case, number, gender
APPR	case
APPRART	case, number, gender
APPO	case
ART	case, number, gender
NN	case, number, gender
NE	case, number, gender
PDS	case, number, gender
PDAT	case, number, gender
PIS	case, number, gender
PIAT	case, number, gender
PIDAT	case, number, gender
PPER	case, number, gender, person
PPOSS	case, number, gender
PPOSAT	case, number, gender

Part-of-speech	**Attributes**
PRELS	case, number, gender
PRELAT	case, number, gender
PRF	case, number, gender, person
PWS	case, number, gender
PWAT	case, number, gender
VAFIN	person, number, mood, tense
VAIMP	number
VMFIN	person, number, mood, tense
VVFIN	person, number, mood, tense
VVIMP	number

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 4.

Table 4: 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 5 lists example tranformations with forms taken from TüBa-D/Z.

Table 5: 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. Each token has an attribute *tf* that marks the field that the token is in.

From the perspective of the TüBa-D/Z, the topological field annotations in TüBa-D/DP are obtained by projecting the topological field node that dominates a given token within the clause onto the token (Kok and Hinrichs 2016; Pütz 2019). If a token is not dominated by a topological field node in its clause, it obtains the special *UNK* label. Figure 1 shows an example of such a projection of topological field nodes onto tokens.

This field projection is only used during the preparation of the training data for the topological field prediction model. During automatic annotation of TüBa-D/DP, the topological fields are directly predicted at the token level.

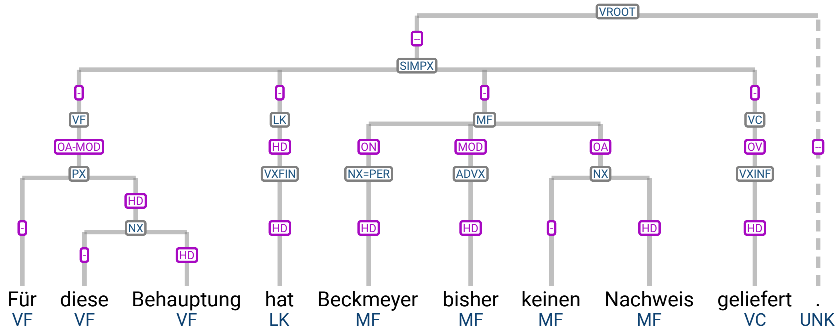

Figure 1: Projection of topological field nodes onto tokens.

3 Annotation tools

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

- **Tokenization**:
  - Wikipedia, Common Crawl: SoMaJo (Proisl and Uhrig 2016)
  - taz: TüPP-D/Z tokenizer (Ule 2004)
- **Part-of-speech tags**: sticker
- **Topological fields**: sticker (Kok and Hinrichs 2016)
- **Dependency relations**:
  - Europarl, Wikipedia, taz: dpar (Kok and Hinrichs 2016)
  - Common Crawl: sticker
- **Morphology**: Marmot (Mueller, Schmid, and Schütze 2013)
- **Lemmas**:
  - Lemmatization: Lemming (Müller et al. 2015)
  - Postprocessing: Ohnomore

References

Buchholz, Sabine, and Erwin Marsi. 2006. "CoNLL-X Shared Task on Multilingual Dependency Parsing." In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, 149–64. Association for Computational Linguistics.

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.

Mueller, Thomas, Helmut Schmid, and Hinrich Schütze. 2013. "Efficient Higher-Order CRFs for Morphological Tagging." In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 322–32. Seattle, Washington, USA: Association for Computational Linguistics. https://www.aclweb.org/anthology/D13-1032.

Müller, Thomas, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. "Joint Lemmatization and Morphological Tagging with Lemming." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2268–74.

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.

Pütz, Sebastian. 2019. "Enriching Topological Field Tagging with Clause Information."

Telljohann, Heike, Erhard W Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2006. "Stylebook for the Tübingen Treebank of Written German (TüBa-d/Z)." In *Seminar Fur Sprachwissenschaft, Universitat Tubingen, Tubingen, Germany.*

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.