Unifying the Treatment of Preposition-Determiner Contractions in German Universal Dependencies Treebanks

Stefan Grünewald[1,2] **Annemarie Friedrich**[2]
[1]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart
[2]Bosch Center for Artificial Intelligence, Renningen, Germany
stefan.gruenewald|annemarie.friedrich@de.bosch.com

Abstract

HDT-UD, the largest German UD treebank, as well as the German-LIT treebank, currently do not analyze preposition-determiner contractions such as *zum* (= *zu dem*, "*to the*") as multi-word tokens, which is inconsistent both with UD guidelines as well as other German UD corpora (GSD and PUD). In this paper, we show that harmonizing corpora with regard to this highly frequent phenomenon using a lookup-table leads to a considerable increase in automatic parsing performance.

1 Introduction

Universal Dependencies (UD) are a cross-linguistic dependency grammar framework driven by a large-scale multi-lingual community effort (de Marneffe et al., 2014). In general, UD prioritizes relations between content words. The treatment of function words, being a rather language-specific issue, is to date sometimes inconsistent even across the treebanks of a single language. Function words including prepositions or negation words are often contracted with other words, which requires to decide whether to keep them as a fused unit or split them up during tokenization – a non-trivial problem.[1]

The German language allows to fuse certain combinations of preposition+determiner into a single token, resulting in what UD refers to as a *multiword token*, i.e., a single token that contains more than one *syntactic word*. Examples include *zum* (= *zu dem, "to the"*) and *ins* (= *in das, "into the"*). These constructions can even be regarded as lexicalized, i.e., as belonging to the inventory of the language's lexicon (Lehmann, 2002). Expanding such contractions into several tokens does not depend on the context, hence, treating them as multi-word tokens is rather straightforward. The only caveat is assigning the correct morphological features to the determiner, but these can easily be retrieved from the head noun.

The current UD annotation guidelines for German[2] suggest treating preposition-determiner contractions as multi-word tokens in the way outlined above, and their treatment is implemented accordingly in the German GSD treebank (292k tokens) as well as in German PUD (21k tokens). However, HDT-UD (Borges Völker et al., 2019), the largest German UD treebank (with 190k sentences and 3.4 million tokens currently the largest available UD treebank overall), as well as the small German-LIT treebank (40k tokens, Salomoni (2017)), do not expand these tokens. This may lead to inconsistency-based parsing errors in cross-treebank experiments or when training a parser on several treebanks.

In this paper, we analyze the extent of the problem, finding that in HDT-UD, 25% of all sentences contain such contractions. Our contributions are (i) the development of a simple lookup-based script for splitting German preposition-determiner contractions into several tokens, and (ii) a set of experiments showing that on the relevant sentences, the increased consistency leads to an increase in parsing accuracy by up to 0.8 points in terms of LAS F1. Hence, this paper constitutes a case study of the improvements we can expect from careful linguistic data analysis and corpus harmonization. We contribute our conversion script, as well as the fixed versions of the HDT-UD and German-LIT corpora, for the next UD release.

[1]See, e.g., https://github.com/UniversalDependencies/docs/issues/641

[2]https://universaldependencies.org/de/index.html

Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020), pages 94–98
Barcelona, Spain (Online), December 13, 2020

Contraction	Expansion	count	% sents
im	in dem	26236	12.8
am	an dem	7764	3.9
zum	zu dem	7584	3.9
zur	zu der	6149	3.1
vom	von dem	3404	1.8
beim	bei dem	2795	1.4
ins	in das	1422	0.7
fürs	für das	233	0.1
ans	an das	160	0.1
übers	über das	147	0.1
TOTAL		56150	25.0

(a) German-HDT-UD

Contraction	Expansion	count	% sents
im	in dem	89	4.2
zur	zu der	44	2.0
zum	zu dem	27	1.4
vom	von dem	17	0.9
am	an dem	17	0.9
ins	in das	8	0.4
aufs	auf das	7	0.3
beim	bei dem	5	0.3
fürs	für das	5	0.3
beym	bey dem	3	0.1
TOTAL		222	9.8

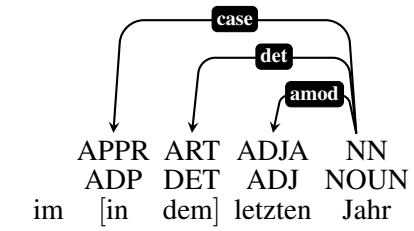
(b) German-LIT

Table 1: The 10 most common contractions in two German UD corpora, as well as the total count. (The computation of the TOTAL row considers the fact that one sentence may contain several contractions.)

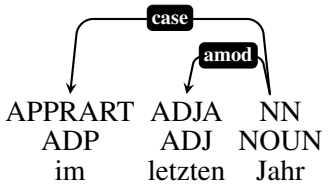
2 Preposition–Determiner Contractions in German UD corpora

The German HDT-UD treebank (Borges Völker et al., 2019) has been automatically converted from the Hamburg Dependency Treebank (HDT, Foth et al. (2014)) using a tree transducer. The text data stems from the German technical website heise.de, which contains, among others, reports about new software and hardware as well as technology-related politics. HDT uses its own dependency annotation scheme (Foth, 2006), in which, in contrast to UD, relations are headed by function words. Preposition-determiner contractions are simply marked as prepositions and indicate their complement (i.e., what would be a preposition's head noun in UD) using the relation PN.

Table 1(a) reports corpus statistics for the occurrences of the most frequent preposition-determiner contractions in HDT-UD. Contractions occur in 25% of all sentences, showing that treating the phenomenon in a consistent way is non-negligible. Some contractions, including *im* (= *in dem*, "*in the*"), *am* (= *an dem*, "*at the*") and *zum* (= *zu dem*, "*to the*") are extremely common; others are more colloquial and rarer in written German (e.g., *übers* = *über das*, "*over the*").



(a) Expanded contraction (according to guidelines).



(b) HDT-UD / German-LIT original version.

Figure 1: Annotation of contractions in the phrase *im letzten Jahr* (*"in the last year"*).

Figure 1 illustrates the two different ways of treating preposition-determiner contractions in German UD corpora. Version (a), introducing two trace-like tokens, is suggested by the official guidelines, and applied in GSD and German-PUD. The commonly used Stanza tokenizer (Qi et al., 2020) also employs this strategy. HDT-UD currently applies a single-token analysis (b). The contractions have their own language-specific XPOS tag `APPRART` ("preposition with article") taken from the STTS tagset (Schiller et al., 1995). Their UPOS tag in HDT is `ADP`, which further shows that the single-token analysis is inadequate: `ADP` stands for adpositions (a cover term for prepositions and postpositions), but the contractions also include a determiner in addition to a preposition. This information is lost in the single-token analysis, which attaches the contraction to its head via *case*, circumventing *det* altogether.

The same issue exists in the small German-LIT treebank (Salomoni, 2017), which consists of short fragments of 18th century literary essays about aesthetical issues by Schlegel and Novalis, written in the then-young Hochdeutsch (modern German). Corpus statistics for German-LIT are given in Table 1(b).

For each token with index i: if the token has the APPRART XPOS tag (and no exception applies*):

1. Increase the indices of all tokens after index i by 1.
2. Insert the contraction as a new multiword token with index $(i, i+1)$
3. Replace the contraction at i with two new tokens: a preposition and a determiner as specified by the lookup table.
4. Attach the preposition and determiner to the contractions's head via the *case* and *det* relations, respectively.
5. Copy the contraction's head's *Case* feature to the preposition and its *Case*, *Number*, *Gender* features to the determiner.

*Exceptions include tokens that are clearly incorrectly tagged and tokens attached via the *reparandum* relation (i.e., disfluencies).

```
1-2 Im      _     _      _ _
1   In      ADP   APPR   3 case
2   dem     DET   ART    4 det
3   letzten ADJ   ADJA   3 amod
4   Jahr    NOUN  NN     4 obl
5   stieg   VERB  VVFIN  0 root
6   der     DET   ART    6 det
7   Umsatz  NOUN  NN     4 nsubj
```

Example CoNLL-U file (output of algorithm). *"In the last year increased the sales"*

Figure 2: Algorithm for lookup-table based preposition-determiner expansion in German UD.

3 Expanding Contractions

Based on the above discussion, we propose a simple lookup-based method for expanding preposition-determiner contractions in German UD corpora into multi-word tokens in order to make the data consistent with the UD annotation guidelines. Figure 2 shows our algorithm, which operates on files of the CoNLL-U format.[3] Note that we do not make changes for the token *z.* from *z. B. = zum Beispiel* = "*for example*," as well as in several other infrequent special cases resulting from annotation errors.

Since prepositions and determiners are closed word classes and the mapping from contractions to their expansions is unambiguous in German, we use a simple lookup table (see Table 1) for expansion. The full table was constructed by extracting all word forms labeled APPRART from the HDT-UD and LIT corpora and then using the authors' knowledge of German to assign the correct expansions. Morphological features of the syntactic words of the expansion may be ambiguous: for example, *zum* may be expanded into *zu dem*$_{Gender=Neut}$ or *zu dem*$_{Gender=Masc}$. However, we can easily derive the correct features by simply copying over the *Case*, *Number*, *Gender* features from the syntactic head of the contraction. (Note, however, that potential annotation errors are also propagated this way.)

We ensure the correctness of the output by manually inspecting some of the resulting annotated sentences as well as running the official UD validation script.[4]

4 Parser Evaluation

To evaluate how our changes to the corpora affect parsing accuracy, we train the state-of-the-art UDify parser (Kondratyuk and Straka, 2019) on the existing GSD corpus as well as the original and modified versions of the HDT-UD corpus, and report scores on various versions of the test sets. (LIT and PUD only provide test data.) We keep all hyperparameters the same as in the original setup, except that (a) we use the German BERT model by deepset[5] to initialize BERT weights; and (b) we only train on HDT-UD for 25 epochs due to the extremely large size of the corpus. We report results for gold tokens.

Table 2 shows parser performance on sentences containing contractions in the test sets of GSD, HDT-UD, and LIT (the latter two in both their original and modified versions, indicated by *+/-exp*). As can be seen, for the parser that is trained on the GSD corpus (in which contractions are split up), performance is higher on the modified versions of the HDT and LIT corpora than on the original versions, as would be expected due to the unified treatment of contractions. Interestingly, the same holds when training on the original version of HDT-UD. The reason for this is that the original corpus also contains many cases of non-contracted preposition-determiner combinations, enabling the parser to get expanded contractions right as well. Furthermore, scores increase on the fixed test sets because the expansion of contractions leads to more *det* relations, which are generally very easy to predict. However, training on the fixed data is still beneficial, as comparing the last two rows for each of the *+exp* test sets shows.

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

[4] https://github.com/UniversalDependencies/tools/blob/master/validate.py

[5] https://deepset.ai/german-bert

	test					
↓ **train**	GSD_{+exp}	PUD_{+exp}	HDT_{-exp}	HDT_{+exp}	LIT_{-exp}	LIT_{+exp}
GSD_{+exp}	85.78	85.34	85.70	**86.50**	79.76	**80.45**
HDT_{-exp}	79.15	81.32	95.57	**95.72**	76.33	**76.91**
HDT_{+exp}	79.23	81.41	95.45	**95.73**	76.34	**77.11**

Table 2: Parsing performance (**LAS F1**) on test set sentences that contain contractions. *+exp* indicates that contractions are analysed as multi-word tokens in the respective corpus, *-exp* indicates that they are analysed as single tokens.

An analysis by label type confirms that increases in accuracy are mainly caused by *case* and *det*. For example, when training on GSD, LAS F1 of *case* increases from 96.28 (HDT_{-exp}) to 96.60 (HDT_{+exp}); *det* increases from 96.74 (HDT_{-exp}) to 97.72 (HDT_{+exp}). We also observe modest improvements in parsing accuracy on a number of other dependency labels such as *obl*, *nmod*, and *ccomp*, indicating that surrounding syntactic constructions also benefit from the consistent handling of contractions.

Interestingly, our results also show that the parser trained on GSD, despite not having encountered contractions during training, still attaches the vast majority of contractions in HDT_{orig}/LIT_{orig} correctly. We suspect that this may be owed to BERT's ability to generalize from simple prepositions to contractions because of their similar distribution in the pre-training data.

5 Related Work

Discussions within UD community. It is an on-going discussion within the UD community how to best achieve a standardized treatment of tokenization and word tokenization, e.g., how to treat multi-word tokens such as *gonna* (= *going to*) in English. A current proposal[6] suggests breaking up these tokens for formal English as was done in the original Penn TreeBank annotations, but allowing different solutions for informal language such as Twitter posts. We follow this suggestion in some sense as we also create multi-word tokens for highly frequent and lexicalized preposition-determiner contractions in German.

Within the context of Surface-Syntactic Universal Dependencies (SUD, Gerdes et al. (2018)), it has been proposed to treat propositions as heads, moving away from UD's focus on content words and applying distributional criteria for the units instead. In fact, in SUD, the question of how to deal with these contractions is even more pressing: Because the preposition part of the contraction would be the head of a noun, but the determiner part would be a dependent, it is not quite clear what the overall syntactic relation of the contraction to the noun should be if left as one syntactic word.

French amalgames. A similar issue arises in the word segmentation of French *amalgames* (contractions). Here, the situation is slightly more complicated due to the ambiguity of *du/des*, which may occur either as indefinite determiners as in *des enfants jouent* ("(some) kids play"), in partitive constructions such as *je bois du lait* (= *de le*, "I drink (some) milk"), or in possessive constructions such as *la lettre des filles* (*de les*, "the letter of the girls"). Currently, the major French treebanks annotate infinite determiner *des* as DET, while they split the other two cases into two tokens. However, a context-dependent treatment of tokenization bears technical difficulties for automatic processing; a discussion by the treebank maintainers suggests to always split these contractions and annotate the indefinite determiner case using a *fixed* relation between the two components.[7] To date, this does not seem to have been implemented.

6 Conclusion

In this paper, we have carefully analysed the treatment of preposition-determiner contractions in German UD corpora. Harmonizing representations lead to increases in LAS F1 of up to 0.8, indicating that a unified treatment of these frequent construction is essential. We here have presented a case study of how to unify word segmentation for the relatively simple case of German preposition-determiner contractions. Future work includes addressing similar phenomena in more difficult situations such as

[6] https://github.com/UniversalDependencies/docs/issues/641

[7] https://github.com/bguil/UD-French-discussion/issues/1

the context-dependent interpretation of French amalgames or, more generally, finding good guidelines for word segmentation and harmonizing corpora accordingly. The latter is especially tricky for informal genres where segmentation decisions seem to be a continuum.

References

Emanuel Borges Völker, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. HDT-UD: A very large universal dependencies treebank for German. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 46–57, Paris, France, August. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Kilian A. Foth, Arne Köhn, Niels Beuck, and Wolfgang Menzel. 2014. Because size does matter: The hamburg dependency treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2326–2333, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Foth, 2006. *Eine umfassende Constraint-Dependenz-Grammatik des Deutschen*.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or surface-syntactic universal dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium, November. Association for Computational Linguistics.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China, November. Association for Computational Linguistics.

Christian Lehmann, 2002. *New reflections on grammaticalization and lexicalization*, volume 49, pages 1–18. 01.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Alessio Salomoni. 2017. Toward a treebank collecting german aesthetic writings of the late 18th century. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it)*.

Anne Schiller, Simone Teufel, and Christine Thielen. 1995. Guidelines für das tagging deutscher textcorpora mit stts. *Universität Stuttgart, Universität Tübingen, Germany*.