# Unifying Preposition-Determiner Contractions in German UD Treebanks

Stefan Grünewald[1,2] Annemarie Friedrich[1]

[1]Bosch Center for Artificial Intelligence [2]University of Stuttgart, Germany

BOSCH
Invented for life

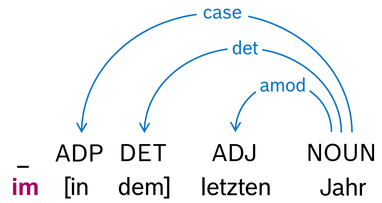
University of Stuttgart
Germany

## OBSERVATIONS:

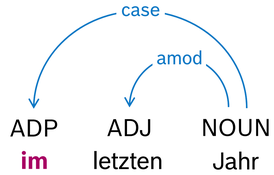
- Preposition-determiner contractions such as *im (=in dem, „in the“)* are treated inconsistently across German UD treebanks.
- UD guidelines suggest a *multiword token analysis* (left) rather than a *single-token analysis* (right).

## CONTRIBUTIONS:

We analyze the extent of this issue in German UD corpora and propose a simple lookup-table based method for automatically converting contractions into the multiword token representation. We show that this results in increased automatic parsing performance.




**Multiword token analysis (GSD and PUD):**
Contractions are split up into preposition and determiner, introducing two trace-like tokens




**Single-token analysis (HDT and LIT):**
Contractions are left as-is and treated like prepositions (*ADP* tag, *case* dependency)

## Corpus Analysis: Contractions in German UD Corpora

**Top 10 contractions in HDT (Borges Völker et al., 2019)**

| 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 |

**Top 10 contractions in LIT (Salomoni, 2017)**

| 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 |

Contractions are highly frequent, occurring in 25% of sentences in the HDT corpus and 10% of sentences in the LIT corpus.

→ Consistent treatment of contractions is non-negligible!

## Lookup-table Based Expansion of Contractions

```
1-2  Im       _      _
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
```

```
1  In   ADP   APPR  3  case
2  dem  DET   ART   4  det
```

- **Rule-based algorithm:** Convert contractions into multi-word tokens, inserting trace-like tokens for preposition and determiner based on a manually constructed lookup table. Inserted traces are attached to the syntactic head; morphological features are copied over from head.
- Exceptions include tokens which are clearly incorrectly tagged and tokens attached via the *reparandum* relation (disfluencies).
- Official UD validation script → check well-formedness of the output.

## Evaluation of Parser Performance

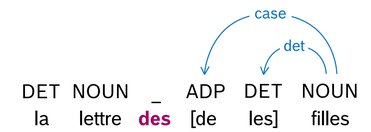
**Parsing performance (LAS F1) on test set sentences containing contractions**

| | 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 |

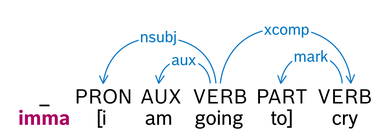
- **Parser**: UDify **(Kondratyuk & Straka, 2019)**
- **Training**: GSD (uses expanded version); original and modified versions of HDT
- **Evaluation**: Test set sentences containing contractions from GSD, PUD, HDT, and LIT
- **Results**: For all parsers, performance is better on the modified versions of HDT and LIT, as expected due to the unified treatment of contractions. Performance on GSD and PUD is better when training on modified HDT.
- **Analysis**: Increases in accuracy are mainly caused by *case* and *det* dependencies. We also observe modest improvements on other dependency labels such as *obl* and *nmod*, indicating that consistent handling of contractions also benefits surrounding constructions.

## Conclusion and Future Work

- We have proposed a simple lookup-table based method for harmonizing the treatment of German contractions.
- Future work includes addressing similar phenomena in other languages, which entails finding consistent guidelines for word segmentation and harmonizing corpora accordingly.




(a) French *amalgames*




(b) Contractions in (informal) English

## References


- **Borges Völker et al. (2019):** HDT-UD: A very large universal dependencies treebank for German. *In: Proceedings of UDW 2019.*
- **Salomoni (2017):** Toward a treebank collecting German aesthetic writings of the late 18th century. *In: Proceedings of CLiC-it 2017.*
- **Kondratyuk & Straka (2019):** 75 languages, 1 model: Parsing universal dependencies universally. *In: Proceedings of EMNLP 2019.*