Subword Relations, Superword Features

Daniel Zeman
Charles University, Faculty of Mathematics and Physics, ÚFAL
Malostranské náměstí 25
CZ-11800 Praha
zeman@ufal.mff.cuni.cz

Relevant UniDive working groups: WG1, WG2

1 Introduction

Universal Dependencies (UD) subscribes to the lexicalist principle, claiming that dependency relations connect *words*, while the process of constructing words by combining smaller units *(morphemes)* is substantially different. Consequently, word-internal structure is normally not shown in UD.[1] Nevertheless, there seems to be some demand (Baldwin et al., 2021) for a UD extension that would allow for showing word-internal structure in a way similar to how inter-word relations are represented. Here are some motivational examples:

- German compounds are written as one word and represented by one tree node in UD. English compounds may be perfectly parallel to the German ones, yet they are typically written as multiple orthographic words. In UD, they are multiple nodes connected by `compound` relations. The parallel structure is not visible in German UD but it could if the compounds were split into multiple tree nodes (Fig. 1).[2] Moreover, other annotation may pertain just to one part of a compound: We may want to annotate the MWE *Rolle spielen* "to play a role" in the compound *Hauptrolle spielen* "to play the main role".

- Turkish words may combine several derivational and inflectional steps. Traditional analysis would break them up to *inflection groups* but in UD they are mostly kept together and the internal structure is not visible (unlike Fig. 2).

- Chukchi transitive verbs may incorporate their objects and switch to intransitive inflection (Fig. 3) (Tyers and Mishchenkova, 2020).

- Fieldworkers may prefer morpheme-based analysis when documenting a language; a UD example is the treebank of Beja (Kahane et al., 2021).

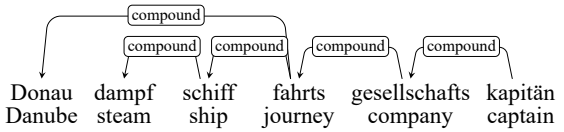

Figure 1: *Donaudampfschifffahrtsgesellschaftskapitän* "Danube steamship company captain"

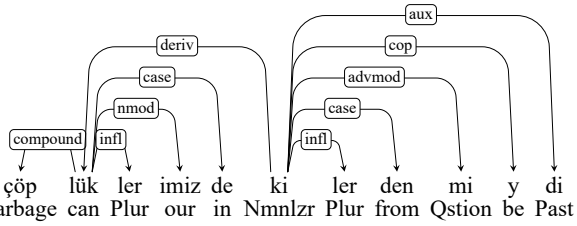

Figure 2: *çöplüklerimizdekilerdenmiydi* "was it from those that were in our garbage cans?"

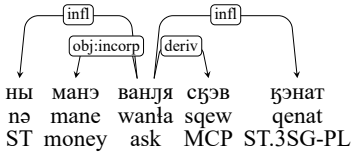

Figure 3: *nəmanewanłasqewqenat* "they constantly asked for money"

Precisely defining a *word* (even a *syntactic word*) cross-linguistically is a difficult task (Haspelmath, 2022 Draft). However, it matters less if we can annotate inter-word and intra-word relations in a similar manner. We propose to work within WG1 (and partially WG2) on an extension of UD that would support such annotation.

2 Subword Relations

As relations between subword units violate the lexicalist principle, they cannot be part of a regular

[1] Except for the optional `MSeg` and `MGloss` attributes in the MISC column of some treebanks, which can at least hint at the morphemic composition of a word.

[2] In fact, compounds are a gray zone. While most UD languages do not split them, they are split in Sanskrit UD, as such analysis is traditional in Sanskrit linguistics.

```
# text = Er spielt die Hauptrolle im Haus.
# text_en = He plays the main role in the house.
1   Er          er          PRON  _ Case=Nom|PronType=Prs 2 nsubj _ _
2   spielt      spielen     VERB  _ Mood=Ind|VerbForm=Fin 0 root  _ _
3   die         der         DET   _ Case=Acc|PronType=Art 4 det   _ _
4-6 Hauptrolle  _           _     _ _                       _ _     _ _
4   Hauptrolle  Hauptrolle  NOUN  _ Case=Acc|Number=Sing  2 obj   _ _
5   haupt       haupt       ADJ   _ Degree=Pos            6 amod  _ _
6   Rolle       Rolle       NOUN  _ Case=Acc|Number=Sing  4 wroot _ _
7-8 im          _           _     _ _                       _ _     _ _
7   in          in          ADP   _ _                       9 case  _ _
8   dem         der         DET   _ Case=Dat|PronType=Art 9 det   _ _
9   Haus        Haus        NOUN  _ Case=Dat|Number=Sing  2 obl   _ SpaceAfter=No
10  .           .           PUNCT _ _                       2 punct _ _
```

Figure 4: CoNLL-U with subword relations.

UD treebank under the current guidelines; they have to be an extension that stands outside UD proper. Nevertheless, the file format should retain low-level compatibility with CoNLL-U so that existing tools can still process it. So, while new relation labels are conceivable, there should be no new line types beyond the existing 5 (comment, multiword token, node, empty node, empty line). There may be extra columns for readability (CoNLL-U Plus[3]) but it should be possible to collapse them into MISC attributes if needed.

Ideally, the format should accommodate normal UD treebank plus additional subword annotation and there should be a script that throws away the extra relations and extracts the regular UD treebank. If a word is decomposed, the relations between its parts should probably form a tree (⇒ single root). The annotation of the root morpheme will differ from the annotation of the whole word, so we need nodes for both.[4] Multiword token lines must be used to indicate the mapping of the nodes to surface tokens (Fig. 4).

3 Superword Features

Conversely, we may want to assign word-like annotation to a multiword expression. For example, a MWE functions like an adverb although its member words are not adverbs. Some treebanks already mark this with `MWEPOS=ADV` (or `ExtPos`) in MISC. Similarly, for German verbs with separable prefixes (e.g. *ein|steigen* "get on"), we may want to indicate the lemma that describes the two parts together. We may also want to add morphological features to sets of words, e.g., `Tense=Fut` for periphrastic future (composed of words that are not future themselves).

The MWE does not have to be linearly contiguous, so we cannot abuse multiword token lines for this purpose. MWEs tend to be catenas,[5] suggesting that the MISC column of the head node could hold such annotations. They are not complete subtrees though: in *I have come home*, the head of the periphrastic verb form *have come* is *come*, but we want to exclude the other dependents (*I* and *home*) from the annotation of the verbal features. We thus need a MISC attribute with the IDs of the nodes that are included in the MWE, e.g., `MWSpan=1-3,5`.

Multiple MWEs could have their annotation placed at the same head node, meaning that we have to use numeric ids to mark MISC attributes that pertain to the same MWE. For example, in *He has played the main role in the process*, we could annotate `MWSpan[1]=2-3 | MWLemma[1]=play | MWUPOS[1]=VERB | MWAspect[1]=Perf` and

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

[4] As one of the reviewers noted, this has drawbacks, too. Parallelism between languages will be somewhat spoiled, as German *Hauptrolle* will now have three nodes, while English *main role* will have only two. Alternatively, the word-level morphological annotation could be stored for the morphemes spanning the word in a similar manner to what we propose for superword features in the next section.

[5] Even catena is probably not always granted. Grouping auxiliaries without the main verb would be a problem, although one may argue that this can be left for SUD to deal with. But coordination may complicate things. In *The food has been cooked and eaten*, one may want to combine the auxiliaries not only with *cooked* but also with *eaten*. Maybe we can say that this would be a catena in the enhanced dependency graph.

`MWSpan[2]=2-3,6` | `MWLemma[2]=play role` | `MWUPOS[2]=VERB` | `MWAspect[2]=Perf`. Essentially, what we are looking at is a constituent-oriented analysis combined with dependencies, although 'constituents' in this sense are not linearly contiguous spans of words.

Acknowledgements

This work was supported by the grants 20-16819X (LUSyD) of the Czech Science Foundation; and LM2023062 (LINDAT/CLARIAH-CZ) of the Ministry of Education, Youth, and Sports of the Czech Republic.

References

Timothy Baldwin, William Croft, Joakim Nivre, and Agata Savary. 2021. Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics (Dagstuhl Seminar 21351). *Dagstuhl Reports*, 11(7):89–138.

Martin Haspelmath. 2022 Draft. Defining the word.

Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 2021. A morph-based and a word-based treebank for Beja. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 48–60, Sofia, Bulgaria. Association for Computational Linguistics.

Francis Tyers and Karina Mishchenkova. 2020. Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 195–204, Barcelona, Spain (Online). Association for Computational Linguistics.