Word Segmentation in Universal Dependencies

Kilian Evang
Heinrich Heine University Düsseldorf
Germany
evang@hhu.de

Daniel Zeman
Charles University, Prague
Czechia
zeman@ufal.mff.cuni.cz

Relevant UniDive working groups: WG2, WG1

1 Introduction

Morphosyntactic annotation in the Universal Dependencies framework (UD) relies on the notion of word (de Marneffe et al., 2021). As they point out, this notion is challenging to define in a cross-linguistically consistent manner. In the context of annotating multiword expressions, the question of what a word is also affects the question of what a multiword expression is (Savary et al., 2023). In this abstract, we summarize the preliminary results of a survey carried out among UD treebank maintainers aiming to uncover to what extent the definitions of "word" employed in different UD languages and treebanks are consistent, and what would be required to unify them further. Because there is no universally accepted definition of "word", and the definitions used in UD treebanks are rarely made explicit, we used Haspelmath (2023)'s operational definition of "word" as a point of reference. This was certainly not the only option, as linguists have been trying to define word since (at least) Bloomfield in 1930s. But Haspelmath claims to have addressed shortcomings of the previous definitions and we decided to confront his approach with corpus data. Our survey consisted of 15 questions, most of them soliciting treebank examples of individual types of words (possibly with comments). We received responses for 40 languages from 12 families. Our analysis reveals that the main challenges lie in a cross-linguistically consistent treatment of clitics and compounds.

2 Tokens and Words in UD

UD distinguishes between (orthographic) tokens and (morphosyntactic) words. For languages whose writing systems mark word boundaries with whitespace, this usually defines the segmentation into tokens. Words are the nodes of the syntactic dependency graph. Divergences between the notions of token and word occur when an orthographic token is subdivided into two or more words (multiword tokens) or two or more orthographic tokens form a single word (multitoken word). In the following, we only focus on UD words and to what extent they are consistently defined across UD treebanks.

3 Types of Words

Haspelmath's definition of word is disjunctive, distinguishing four types. In this section, we briefly discuss for each type its representation in the surveyed UD treebanks. We also discuss cases of UD words that are not words according to Haspelmath, and vice versa.

Free morphs Free morphs are single morphs that can occur on their own, e.g., as a response to a question. They occur in all languages, at minimum, as interjections. Free morphs seem to be consistently treated as UD words in the surveyed treebanks.

- *muž* "man" (Czech)
- *land* "country" (Afrikaans)

Roots (with affixes) Roots are contentful morphs (denoting an object, property, or action) that can occur without any other contentful morph (but not all roots are free morphs because they may require affixes). The surveyed treebanks seem to be consistent in treating roots (together with all their required affixes if any, plus potentially nonrequired affixes) as words. A possible exception is the grey area of morphs whose status as affix vs. clitic is unclear, see below. In some languages, roots do not typically have any required affixes (e.g., Afrikaans, English, Hebrew); in others, (almost) all roots do (e.g., Ancient Greek), and some languages fall somewhere in the middle (e.g., Brazilian Portuguese). We did not discover any instances where a morph is clearly an affix according to Haspelmath but is treated as a word separate from its root in UD.

- *muž-e* "man-GEN.SG" (Czech)
- *bonit-a* "beautiful-FEM.SG" (Portuguese)

Clitics Clitics are morphs that cannot occur without a root (meaning they are not roots themselves) but, unlike affixes, they are not selective about the

category of the root they occur with. This definition covers many morphs that often appear in UD treebanks as words with the parts of speech ADP, DET, PART, PRON, CCONJ, and SCONJ. However, in some cases, the survey revealed difficulties in reliably applying the definition and distinguishing clitics from affixes. The selectivity towards the category of the nearest root requires defining what a root category is and also leaves open the possibility of treating a morph as either a single clitic or a bunch of homonymous affixes. An example is the Czech negation morph *ne* which might be a clitic according to Haspelmath but is treated as an affix (part of larger words) in the UD treebanks.

- *en* "in" (Ancient Greek)
- *li* "interrogative particle" (Bulgarian)

Compounds (with affixes) According to Haspelmath's definition, compounds are strictly adjacent roots. The grammatical and orthographic traditions of many languages disagree with this, treating as (compound) words also certain combinations of roots with linking morphs in between (e.g., Czech, Dutch, Eastern Armenian). UD usually follows this tradition and treats such compounds as words.

- *pół-noc-y* "mid-night-GEN.SG" (Polish)
- *dà-xué-shēng* (lit. *big-study-raw*) "college student" (Chinese)

UD Words that are not words according to Haspelmath The main type of UD word that is not a Haspelmath word is compound with linking elements (see paragraph "Compounds (with affixes)"). In addition, there are some possible clitics that are treated as parts of words in certain UD treebanks (see paragraph "Clitics").

- *Liebe-**s**-brief* "love-GEN-letter" (German)
- ***ne**-zn-ám* (lit. *not-know-1.SG*) "I don't know" (Czech)

Words according to Haspelmath that are not UD words There are two main types. The first are compounds like the English *farm land* – a (compound) word according to Haspelmath, but written with a space and thus treated as two words (linked with the compound relation) in UD. The second type possibly belonging here is contractions (e.g. of a preposition and an article). Haspelmath does not discuss them specifically and it is debatable whether they can be still seen as multiple morphs; if not, then the whole contraction would be a word, but UD often splits it into the components that were contracted.

- *shukutoku daigaku* "Shukutoku University" (Japanese)
- *em* + *a* → *na* "in the.FEM.SG" (Portuguese)

4 Towards a More Cross-lingually Consistent Definition of "Word" in UD

A universal definition of "word" would be applicable across languages and not rely on individual-language grammatical and orthographic traditions. We have to acknowledge that these traditions play an important role in making UD accessible for a broad audience, so it is not realistic to assume that they can be completely abandoned. Yet it is desirable to have a working definition for UD that can be used for guidance especially in situations where the intra-language tradition is not strong or the language is not written at all. Elsewhere, the definition at least increases our understanding how treebanks differ, and encourages explaining such divergences in documentation.

Our survey has revealed that the treatment of free morphs as well as roots with affixes is already mostly consistent in UD, and conforms to Haspelmath's definition. The main areas of work where more consistency could be achieved is clitics, compounds, and contractions. In the following, we give some preliminary recommendations.

Clitics Because Haspelmath's definition of clitic still seems somewhat vague, we have not been able to determine to what extent UD treebanks are consistent with it, or with each other. More precise criteria will have to be developed, especially to clearly demarcate clitics from affixes.

Compounds This is perhaps the area of most divergences between treebanks and Haspelmath, but also inconsistencies within UD as such. Haspelmath's definition creates a divide between compounds with and without linking elements that seems unnatural in many languages. As an alternative way to handle this in UD in a cross-lingually consistent manner, forced splitting of compounds might be an option, allowing only one root per UD word, even when the grammatical and orthographic traditions say otherwise.

Contractions There is an established practice in UD of splitting contractions into their etymological components. A crosslinguistically applicable criterion for when to do this remains to be formulated.

Acknowledgements

We would like to thank the respondents to our survey for their invaluable and thoughtful input. We also thank the three anonymous reviewers for their feedback. The first author's work was supported by the grant 467699802 (MWESemPrE) of the Deutsche Forschungsgemeinschaft. The second author's 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

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Martin Haspelmath. 2023. Defining the word. *WORD*, 69(3):283–297.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023. PARSEME meets Universal Dependencies: Getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology (NEJLT)*, 9.