UD annotation tools and best practices for the creation of morpho-syntactic tree-banks for low-resource languages

Omer Goldman
Bar-Ilan University, Israel
omer.goldman@gmail.com

Abstract

As part of the UniDive shared-task, I am co-leading an effort to amend dependency parsing data to make it independent of word boundary. This will allow natural treatment of many low-resource poly-synthetic language as well as making existing UD tree-banks more parallel across languages. While most of the morpho-syntactic data for the shared-task will come from relatively minor extension of existing data in several languages, we also want to construct new tree-banks directly in a morpho-syntactic form. Being new to tree-bank construction, attending the training school will help me close that gap and be more involved in creating the data for the shared-task.

Keywords: morpho-syntax, typology, linguistic diversity, word boundaries

1. Motivation

Words have long been an essential concept in the definition of treebanks in Universal Dependencies (UD; de Marneffe et al., 2021), since the first stage in their construction is delimiting words in the language at hand. This is done due to the common view in theoretical linguistics of words as the dividing line between syntax, the grammatical module of word combination, and morphology, that is word construction (e.g., Dixon and Aikhenvald, 2002).

However, delimiting syntactically relevant words gets exponentially more complicated the less isolating languages are. Thus, this operation, which is as simple as breaking the text on white spaces for English, is borderline impossible for polysynthetic languages, in which a single word may be composed of several lexemes that have predicate-argument relations. This reflects the fact that despite the presumed role of words in contemporary linguistics, there is no consensus on a coherent cross-lingual definition of words (Haspelmath, 2011).

The project presented here presents the ongoing effort to redefine morpho-syntactic data and tasks to naturally apply to a set of languages that is as universal as possible. In line with previous theoretic (Tesnière, 1959; Anderson, 1992) and practical works (Nivre et al., 2022; Goldman and Tsarfaty, 2022), we suggest defining the content-function boundary to differentiate "morphological" from "syntactic" elements. In our morpho-syntactic data structure, content words are represented as separate nodes on a dependency graph, even if they share a whitespace-separated word, and both function words and morphemes contribute morphology-style features to characterize the nodes.

We will thus avoid (most) theoretical debates on word boundaries, and solve much of the word segmentation inconsistencies that occur in UD, either across languages[1] or across treebanks of the same language.[2] Morpho-syntactic data will be more inclusive towards languages that are currently treated unnaturally, most prominently noun-incorporating languages. Morpho-syntactic models will be able to parse sentences in more languages and enable better cross-lingual studies.

2. The Resource

The resource we intend to create in this project is a set of morpho-syntactic treebanks. In practice, the difference with the existing UD schema is relatively mild. As can be seen in Table 1 for an Engilsh example, we will amend the CoNLL-U file format to include MS-features only for content words. Including all nodes while ignoring the MS-features will result in a regular dependency graph, while ignoring all nodes without MS-features will result in a morpho-syntactic tree for this task.

Given this schema, the annotation process is expected to be substantially different for languages with existing UD treebanks and for languages we would like to annotate anew, mostly polysynthetic languages that will demonstrate the superiority of our schema. For languages with preexisting UD treebanks we will amend the CoNLL-U files to include MS-features only for content words. These features will be annotated mostly automatically from current morphological features and function words. However, for languages with no UD data, the annotation process will have to relay on other sources or start from an absolute *tabula rasa*.

[1] E.g., Japanese is treated as isolating and Korean as agglutinative, although they are very similar typologically.

[2] E.g., the different treebanks for Hebrew segment and attribute different surface forms for clitics.

ID	Form	Lemma	POS	Feats	Head	DepRel	MS-Feats
1	you	you	PRON	Nom,2,Sing	4	nsubj	Nom,2,Sing
2	will	will	AUX	Fin	4	aux	
3	not	not	PART	Neg	4	advmod	
4	go	go	VERB	Inf	0	root	Fin,Ind,Fut,Neg
5	because	because	SCONJ	-	9	mark	
6	you	you	PRON	Nom,2,Sing	9	nsubj	Nom,2,Sing
7	were	be	AUX	Fin,Ind,Past,2,Sing	9	cop	
8	my	my	PRON	Gen,1,Sing	9	nmod:poss	Gen,1,Sing
9	student	student	NOUN	Sing	4	advcl:because	Sing,Ind,Past

Table 1: An example of a CoNLL-U file for the Enlish sentence "you will not go because you were my student". Note that all columns are preserved and only the last column is added, distinguishing content nodes (in black) from function nodes (in red).

3. Training School Benefits

As a first stage we intend to convert some UD treebanks to a morph-syntactic form and write annotation guidelines in the process, this stage should well over and done by the time of the training school. The second phase, that will include building of new treebanks for low-resource languages is the most crucial benefit of the training school.

The infrastructure and tools needed for the collection from scratch of morpho-syntactic data are completely unfamiliar to me and the training school could be a good opportunity to get acquainted with as many of them as possible and to find the ones best suited for the needs of the new format, that deviate from usual dependency parsing data.

Although the annotation process is likely to include researchers better versed in the targeted languages, attending the training school will allow me to better assist with problems with the annotation, with the tools, or with any unclarities in the guidelines.

4. Open Questions

Tools that Handle Features The state of affairs regarding morphological features, i.e., their perceived marginality and absence from many treebanks, is vividly contrasted with the importance given to features, both morphological and morpho-syntactical, in this project. As a results, many tools for annotation, manipulation, or processing of dependency trees completely ignore all features ascribed to the nodes.

Therefore, as an open question in the training school, I would like to explore the tools that are flexible enough to handle features, mostly for annotation and visualization. In case there will be found none that answer the requirements set by this project, I would like to explore the options for adjusting existing tools or even write something new for this purpose.

Treebanks for Polysynthetic Languages One of the desiderata for the shared-task is treebanks in new languages that are poorly treated by the current schema in UD. These treebanks are likely to be annotated as morpho-syntactic from the beginning, but ideally we would like these resources to be available also in a UD form.

To this end I would like to explore the options at hand when applying the UD schema to languages with different definitions of words. I am aware of some works that have done a similar thing for polysynthetic languages (Tyers and Mishchenkova, 2020; Park et al., 2021) or for agglutinative Turkic languages (Washington et al., 2024), but I believe that a more holistic view with our project in mind that may be more benefitial and may identify where the morpho-syntactic approach is most needed rather than try to force these languages into the UD mold.

5. Current State of Affairs

As of April 2024, much progress has already been done with this project. The schema has been defined, first in low resolution by a handful of researchers, then in details with an extended discussion group. In addition, the effort to annotate data for languages with existing UD treebanks is already underway, as well as the write up of provisional guidelines. I expect that by July the annotation of treebanks for new languages, which is the main purpose of the training school for me, will be immanent. Thus, attending the training school will be as beneficial as possible.

6. Bibliographical References

Stephen R Anderson. 1992. *A-morphous morphology*. 62. Cambridge University Press.

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

Robert MW Dixon and Alexandra Y Aikhenvald. 2002. Word: a typological framework. *Word: A cross-linguistic typology*, pages 1–41.

Omer Goldman and Reut Tsarfaty. 2022. Morphology without borders: Clause-level morphology. *Transactions of the Association for Computational Linguistics*, 10:1455–1472.

Martin Haspelmath. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 45(1):31–80.

Joakim Nivre, Ali Basirat, Luise Dürlich, and Adam Moss. 2022. Nucleus Composition in Transition-based Dependency Parsing. *Computational Linguistics*, 48(4):849–886.

Hyunji Hayley Park, Lane Schwartz, and Francis M. Tyers. 2021. Expanding universal dependencies for polysynthetic languages: A case of st. lawrence island yupik. In *Proceedings of the 1st Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, Online. Association for Computational Linguistics.

Lucien Tesnière. 1959. Éléments de syntaxe structurale.

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.

Jonathan Washington, Çağrı Çöltekin, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, Gulnura Jumalieva, Aida Kasieva, Aslı Kuzgun, Büşra Marşan, and Chihiro Taguchi. 2024. Strategies for the annotation of pronominalised locatives in turkic universal dependency treebanks. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD)*. Association for Computational Linguistics.