A morph based and a word based treebank for beja

Rayan Ziane, Sylvain Kahane, Martine Vanhove, Bruno Guillaume

A MORPH-BASED AND A WORD-BASED TREEBANK FOR BEJA

Sylvain Kahane*, Martine Vanhove**, Rayan Ziane***, Bruno Guillaume****

*Modyco, Université Paris Nanterre CNRS **Llacan, CNRS, INALCO & EPHE
***Llacan, CNRS, INALCO, EPHE & Université d'Orléans ****Sémagramme, INRIA Nancy Grand Est

Introduction

- In Universal Dependencies (UD), basic units of annotation are syntactic words:
  - effective on languages with a low morphology
  - problematic with agglutinative languages with a complex morphology
- We propose to tokenize at morph level:
  - consistent with M. Vanhove's original Beja corpus
  - preservation of all original information
  - automatic transformation into UD standard tokenisation

Beja

- Language of the North Cushitic branch of the Afroasiatic phylum
- Mostly spoken in eastern Sudan, also in the southern part of Egypt and northern Eritrea
- 2,000,000 speakers in Sudan
  - the language has no official recognition
  - does not have its own writing system
  - is not written
- Syntactically, Beja is a final-headed language

From IGT to UD

We built the treebank from a glossed corpus segmented into morphs, in five steps:

1. Data format : (eaf, ELAN software file type, to conll) and **2. Annotation tagset conversion :** (LGR to UD).

3. Automatic pre-annotation : Since our tokenization is morphological, much of the annotations related to the links between affixes/clitics and stems were performed automatically by the Grew tool [2].

4. Manual annotation : Our annotation of dependency relations is manually done in the Surface Syntactic Universal Dependencies (SUD) format [1].

5. SUD to UD : The conversion between SUD and UD was done automatically from a Grew rewriting grammar.

Processing workflow

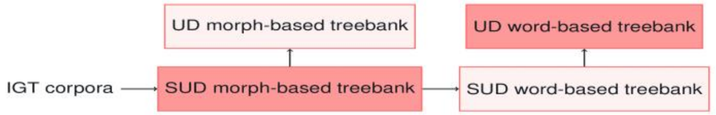


Fig. 1: Treebanks processing workflow

A morph-based treebank

1. Morphological tokenization :

The choice of morphological tokenization is based on the original corpus (figure 2):

- transcribed (T2)
- segmented into word (T3)
- segmented into morphemes (T4)
- glossed (T5, T6)
- translated (T7)
- time-aligned

Using the word level would be a loss of information which goes against our goal of preserving this corpus.

Fig. 2: ELAN screenshot of the source IGT corpus

2. Morphology encoding:

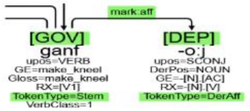


Fig. 3: TokenType feature and "aff" subrel

An organization of information related to morphology processing **in accordance with UD standards**:

- A **TokenType feature** that distinguishes affixes from roots and clitics
- An **"aff" sub-relation**, that specifies that the relation is between a stem and an affix

We consider morphs as if they were words with same function, for instance a nominalizer affix will be treated as a SCONJ with the relation mark (figure 3).

3. A SUD morph-based treebank :

The issue of derivational affixes highlights the direct benefits of a SUD word-based annotation:

- a more detailed syntactic description of the phenomenon
- **an insight in which order the morphs are merged**

This order defines the syntactic category of the concatenated word. NOUN in the schema below takes the category given in the governor's DerPos feature in figure 5.

Fig. 4: UD-style morph-based annotation

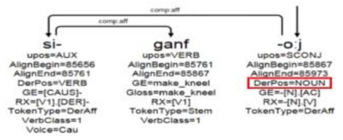


Fig. 5: SUD-style morph-based annotation

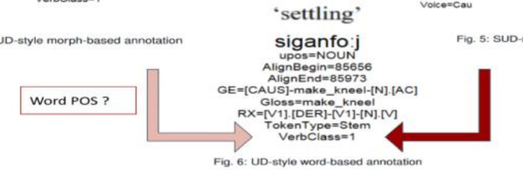
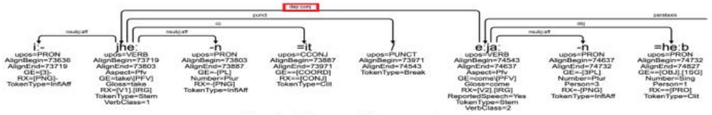


Fig. 6: UD-style word-based annotation

Some constructions in Beja

1. Coordination :

Dependency orientation issue for final-head languages:
As the **conj** relation is forbidden from right to left in UD, we introduced a **dep:conj** relation (figure 7).

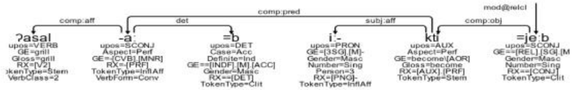


'they took them and they came to me'

Fig. 7: Verbal coordination

2. Relatives clauses :

In addition to the canonical word order for final-head languages, relatives can be constructed with a preposed antecedent, for instance in an aftertought as in figure 8.



'(The man was carrying on his shoulder a lamb.) That he had grilled.'

Fig. 8: SUD style relative clause construction

Conclusions

This treebank of 56 sentences is a small treebank. In its word-based version, it counts 858 tokens against 1157 in its morph-based version. However, it has a complete annotation resulting from the conversion of interlinear glosses and our syntactic manual annotation. The tokens are also time-aligned to the original audio files. The morph to word and SUD to UD conversions were performed automatically by the graph rewriting tool, Grew.
With this treebank we intend to show UD some of these restrictions, with the idea of initiating a debate and enriching the initiative which is in full bloom. Proposing this type of process, which is closer to what field linguists are used to, would allow us to expand the network, the language inventory and the Universal Dependencies project.

References

[1] Kim Gerdes et al. "SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD". In: *Universal Dependencies Workshop 2018*. Brussels, Belgium, Nov. 2018. URL: https://hal.inria.fr/hal-01930614.

[2] Bruno Guillaume. "Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. 2021, pp. 168–175.