A morph-based and a word-based treebank for Beja
Sylvain Kahane*, Martine Vanhove**, Rayan Ziane***, Bruno Guillaume****
*Motycy, Université Paris Nanterre CNRS **Llacs, CNRS,INALCO & EPHE
***Llacs, 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 tokenization

Beja

- Language of the North Cushitic branch of the Afroasiatic phyllum
- 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 target conversion: (IGR to UD).

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

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

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

Processing workflow

UD morph-based treebank
UD word-based treebank
IGT corpora
SUD morph-based treebank
SUD word-based treebank
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 (T3)
- segmented into word (T3)
- segmented into morphemes (T4)
- glossed (T5, T6)
- translated (T7)
- time-aligned

1. Coordinate

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).

Some constructions in Beja

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

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, Grow. 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