Universal Dependencies for Historical Languages

ICHL 26 - Exploiting standardized cross-linguistic data in historical linguistics

Silvia Luraghi, Chiara Zanchi, Erica Biagetti, Luca Brigada Villa
(University of Pavia)

8 September, 2023

UNIVERSITÀ
DI PAVIA

Outline of this presentation

What is Universal Dependencies

What is a treebank

Treebanks:

- collection of morphosyntactically annotated sentence trees
- the appearance of the tree structures may vary depending on the framework used:
  - constituency treebanks
  - dependency treebanks

Dependency treebanks

The importance of words:

- dependency treebanks rely on the dependency grammar
- words are connected to each other in order to form a tree in which each node is represented by a word
- each word has a unique parent node except for the root node
- different annotation styles (Prague style, Universal Dependencies, Surface syntax UD...)

Universal Dependencies

Universal Dependencies:

- cross-linguistic initiative for standardized and shared dependency annotation guidelines and comparable treebanks
- initiated in 2014 with a global collaboration of researchers
- aims to provide a common framework for dependency parsing for a wide range of NLP tasks
- currently includes treebanks for over 100 languages, publicly available for research and development
- it also includes treebanks of some historical languages. Most of them were (semi-)automatically converted from already existing resources

Format

UD treebanks are formatted in **CoNLL-U**:

- each line represents a *syntactic word*
- sentences are separated by a blank line
- each line consists of ten tab-separated fields to store the annotation (`id`, `form`, `lemma`, `upos`, `xpos`, `feats`, `head`, `deprel`, `deps`, `misc`)

Example of a token in CoNLL-U format

```
26 assistant assistant NOUN NN Number=Sing 29 nsubj 29:nsubj _
```

Reasons for starting UD4HL

Why starting this discussion group

We found several reasons to start this discussion group:

- only few historical languages (HLs) are included in UD
- limited number of researchers that actively use treebanks for HLs
- conversion issues:
  - many of the treebanks of HLs come from a conversion from other formats
  - this brings errors in the annotations of the converted resources
  - it is difficult to automatize the conversion from other formalisms
- consistency issues:
  - some comparable constructions of HLs are annotated in different ways
  - At least in the Indo-European family, we find specific constructions that do not find straightforward equivalents in modern languages (so, they lack targeted annotation guidelines)
  - both issue are a consequence of the lack of speakers of HLs, which makes it difficult for annotators to understand what is the best way to annotate a certain structure, especially if it is not present in the modern languages they know

Why starting this discussion group

- HL specific issues:
  - broken sentences
  - necessity of keeping multiple manuscript variants
- we believe that opening a discussion group like UD4HL could improve:
  - the communication among scholars that work with historical languages to produce better resources and guidelines
  - the already available resources, by promoting a process of review and corrrection
  - the communication between the communities of historical language researchers and UD developers

Expected results

As outcome of the discussion we plan to carry on within this group, we expect:

- to start a process of correction and standardization of the already existing resources
- to write better guidelines for researchers that will annotate treebanks for historical languages in the future

Some clarifications

Our initiative is not intended to replace or compete with the existing UD community and the ongoing discussions.

Rather, we aim to establish a bridge between researchers who specialize in historical languages and the UD community. We believe that this collaboration will facilitate mutual learning and foster a deeper understanding of linguistic diversity.

Our goal is to complement and enhance the ongoing efforts of the UD community, not to replace them.

UD4HL discussion group

Composition

Very diverse composition in terms of provenience and backgrounds.

Some numbers:

- **75** researchers involved
  - representing more than **50** different institutions and universities
  - from **25** different countries
  - experts of **30** different historical languages or families

Backgrounds

Concerning the familiarity with treebanks, the backgrounds of the people of this group are very diverse.

Figure: Answers to the question *Do you have any experience with UD treebanks?*

Backgrounds

And also regarding the knowledge of historical languages.

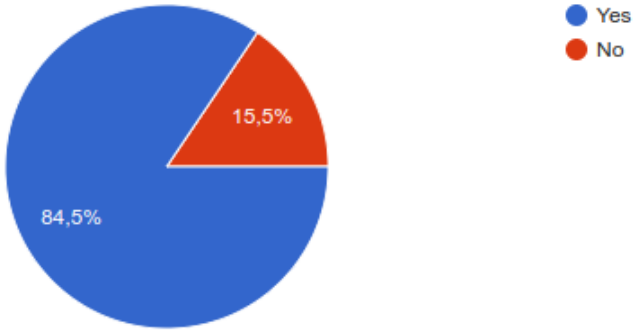
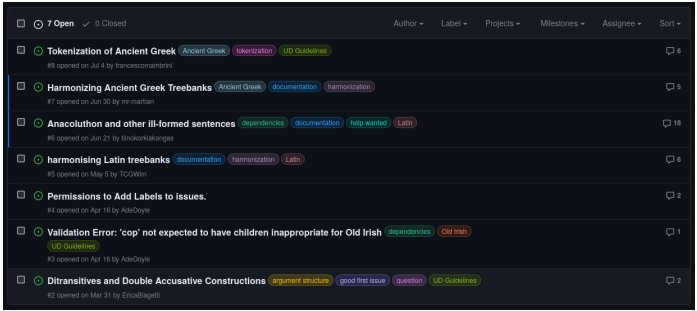


Figure: Answer to the question *Are you an expert of a historical language?*

Work done so far

How we contucted our discussion

We set up a GitHub repository in which we started some threads in the issues section.



You can read the discussion checking out the issues section of the repo:
https://github.com/unipv-larl/UD4HL

How we conducted our discussion

A part from the discussion in the issues section, we met three times to discuss open issues.
Meetings were structured as follows:

- introduction to point out the latest issues open
- three talks given by some members of the group to present a specific issue or a challenge regarding the annotation of the language of their interest
- some time for question, discussion

Topics of presentations during our meetings

Topics covered in the presentations:

- annotation issues:
  - Annotation of double accusative constructions in Ancient Greek, Latin and Sanskrit (with a focus on iobj)
  - Problems regarding the `cop` deprel in Old Irish
- harmonization and conversion issues:
  - process of harmonization of Latin treebanks in UD
  - process of conversion of the PROIEL Classical Armenian treebank into UD format
  - challenges of building a treebank for Old English starting from a constituency treebank
  - methods and tools to convert Ancient Greek corpora in UD

UD4HL: A valuable resource for discussion

We think we all benefit from the discussions during the meetings. Having a discussion group so diverse in terms of backgrounds of the members allowed everyone to learn and exchange ideas.

Often projects are carried on by small groups of researchers and there are few opportunities to get feedbacks from a larger community. We believe that, in particular for resources like UD treebanks that aim to be consistent and comparable, sharing ideas with other experts in Historical Languages, annotation of treebanks, tools for conversion is essential.

After discussing issues $\rightarrow$ moving the discussion to the UD community

UD4HL: A valuable resource for discussion

Communities of annotators of UD treebanks are active, but often lack of the presence of experts of historical languages.

If you want to join us, don't hesitate to ask, open issues in the repository, take part in the discussion!

Next meeting: TBD, around mid October

Thank you for your attention!

- luraghi@unipv.it
- chiara.zanchi01@unipv.it
- erica.biagetti@unipv.it
- luca.brigadavilla@unibg.it