2 Universal Dependency for Historical Languages (UD4HL): Towards Standardized Syntactic Data for Historical Languages

Luca Brigada Villa[1,2], Erica Biagetti[2], Chiara Zanchi[2], Silvia Luraghi[2]

[1] University of Bergamo
[2] University of Pavia

Over the past few decades, historical linguistic research has been enriched with the creation of treebanks for several ancient languages. Most developers have adopted the same annotation schemes employed for treebanks of modern languages, often choosing between the two de facto standards of the Penn Treebank phrase-structure format and the Prague Dependency Treebank (PDT) format. The PROIEL scheme (https://dev.syntacticus.org/proiel.html), which integrates Dependency Grammar with elements of Lexical Functional Grammar and was originally designed for a parallel treebank of translations of the Gospels in old Indo-European (IE) languages, has been applied to several other texts and is nowadays regarded as a further standard for the annotation of historical IE languages (Eckhoff et al. 2018). The multiplication of projects has led to an ever-growing number of historical treebanks that are incompatible with one another. As a result, new treebanks are created for languages that have already been annotated, but according to a different formalism from the one adopted by the authors. Recently, the annotation scheme designed within the Universal Dependency initiative (UD; Nivre et al. 2016, https://universaldependencies.org) has established itself as the standard for dependency annotation. As it favors comparative research, several constituency and dependency treebanks of ancient languages have been converted to UD (notably, we have no knowledge of dependency treebanks being converted to the Penn scheme), and others are now being developed according to this scheme. Yet the achievement of a comparable dataset for historical languages is still hampered by problems related to: a) coverage and balance of each sub-corpus, b) errors caused by the conversion process, and c) the absence of sufficiently clear and adequate guidelines for the annotation of historical languages.

In this paper, we present the state of the art, some issues and possible solutions to obtain corpora as representative as possible of historical languages. In order not to contribute to the flourishing of individual initiatives, we will open a UD working group dedicated to the annotation of such languages in UD: Universal Dependency for Historical Languages (UD4HL). In this group, we plan to address the following issues with the community. First, tools designed to convert the treebanks to the UD format, such as UDConverter (https://github.com/thorunna/UDConverter) and proiel-cli (https://github.com/proiel/proiel-cli), need to be further improved to produce cleaner outputs. Second, we aim to stimulate a revision process of both converted and native UD treebanks that tackles one construction type at a time (cf. Brigada Villa et al. 2022 , Biagetti et al. 2022): this will make it possible to fix errors caused by the conversion and to provide accurate and consistent guidelines for the annotation of new texts. Finally, the conllu format employed by UD features a MISC (miscellaneous) field that can be enriched with information that is not strictly syntactic but useful for studies on the syntax of historical languages, and is currently underexploited. We propose to add various types of information, such as e.g., metrical information for poetic texts or semantic information regarding the animacy of verbal arguments (PROIEL that had such information in its native format, but this has not been included in the UD converted treebanks). Findings and conclusions reached within the working group will be presented at the conference.

Biagetti, Erica, Chiara Zanchi and Francesco Mambrini. Universal Homeric Dependencies? Towards a complete and updated UD treebank of the Homeric poems. Delbrück Symposium on Indo-European Syntax. Università di Verona, November 9-12 2022.

Brigada Villa, Luca, Erica Biagetti and Chiara Zanchi (2022). Annotating “Absolute” Preverbs in the Homeric and Vedic Treebanks. Proceeding of the Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022; Marseille, June 25, 2022).

Eckhoff, Hanne, Bech, Kristin, Eide, Kristine, Bouma, Gerlof, Haug, Dag T. T., Haugen, Odd E. & Jøhndal, Marius. 2018b. The PROIEL treebank family: A standard for early attestations of Indo-European languages. Language Resources and Evaluation 52(1): 29–65.

Nivre, Joakim, De Marneffe, Marie-Catherine, Ginter, Filip, Goldberg, Yoav, Hajic, Jan, Manning, Christopher D., McDonald, Ryan, et al. 2016. Universal Dependencies V1: A Multilingual Treebank Collection. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), 1659–66.