Consolidating the UD Annotation for Armenian

Keywords: Armenian; morphosyntactic annotation; Universal Dependencies

The current release of Universal Dependencies database (v.2.15) includes treebanks of three literary varieties of Armenian, Modern Eastern Armenian (MEA; "UD Armenian BSUT"; "UD Armenian ArmTDP"), Modern Western Armenian (MWA; "UD Western Armenian ArmTDP"), and Classical Armenian (CA; "UD Classical Armenian CAVaL"). Although these varieties have many shared morphosyntactic features, they also show significant differences, conditioned by diachronic and/or dialectal divergence. Besides, the treebanks of the modern varieties, on the one hand, and the one of Classical Armenian, on the other hand, have been developed by different teams of annotators, who do not always follow the same approach to the UD annotation guidelines. These two factors result in only partial compatibility of annotation across the treebanks.

The paper offers a systematic revision of discrepancies in the tagsets and principles of their application. The goal of this analysis is to identify which differences of annotation can be harmonized, and which are conditioned by genuinely dissimilar grammatical features.

The presentation will include a complete chart of correspondences between the tagsets for POS (addressing, where relevant, diverging closed lists of lemmas) and grammatical features with their values. Case studies will illustrate major types of differences, two of which are signalled below.

1) The MEA treebanks offer context-dependent interpretations for values of some morphological features, whereas the CA one follows the morphological principle, which requires to apply the same tag to a specific form. Thus, the feature of "Animacy" of the MEA treebanks has two values "Hum" (human) and "Nhum" (non-human), which are assigned to nouns based on the semantic interpretation of participants that they express. MEA does have the morphosyntactic expression of the contrast between human and non-human direct objects, which are flagged by the dative and nominative-accusative cases, respectively; this grammatical information can be retrieved by a combination of tags "Case=Dat" + "obj" and "Case=Nom" + "obj", respectively. However, the MEA treebanks apply the values of the "Animacy" feature without regard to this pattern, so that one finds combinations like "Animacy=Nhum|Case=Dat" + "obj". By contrast, the CA treebank assigns the values "Anim" (animate) and "Inan" (inanimate) of "Animacy" to pronouns and deternimers that have parallel sets of forms corresponding to these values. Similar issues concern the features of "Aspect" and "Voice".

2) In the MEA treebanks, nouns with enclitic articles and verbs with proclitic negation are treated as single-token words carrying relevant features. By contrast, in the CA treebank, these structures are annotated as groups of separate tokens spelled without a space, and relevant features are distributed among the constituents.

The discussion of results can be helpful for developing hybrid multi-variant parsing models based on UD (see Vidal-Gorène et al. 2024) as well as for developing UD treebanks for other varieties of Armenian such as Middle Armenian, or modern Armenian dialects.

References

Chahan Vidal-Gorène, Nadi Tomeh, and Victoria Khurshudyan. 2024. Cross-Dialectal Transfer and Zero-Shot Learning for Armenian Varieties: A Comparative Analysis of

RNNs, Transformers and LLMs. In Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 438–449, Miami, USA. Association for Computational Linguistics.