

## Towards a UD treebank for Kyrgyz

Aida Kasieva, Gulnura Dzhumalieva, Anna Thompson, Jonathan Washington

In this talk, we present UD annotation work done separately by our two teams, the status of that work, and our plans to combine efforts. Currently there are over 150 annotated sentences, comprising over 1000 tokens, spanning the genres of proverbs and literature. We discuss some of the interesting things that can be learned by working with these genres, and some specific questions of annotation that have come up in our annotation work, including copula tokenisation, analysis of fragments, direction of parataxis attachment, how to deal with certain types of "small" words, how to handle certain non-finite verb constructions, and the level of morphological analysis in relation to POS identification. These are issues with the lowest levels of inter-annotator consistency. We are interested in feedback and discussion on all topics related to annotation, as well as guidance for best practices in submitting these combined "micro" corpora to the UD project.

Keywords: Kyrgyz language, Universal Dependencies, Turkic languages, treebanks

Presentation link:

<https://jonorthwash.github.io/2023-KyrgyzUD-presentation/presentation.html>