UD-Tatar NMCTT Treebank: Issues in annotation across Turkic UD

This report presents UD Tatar NMCTT Treebank and raises several issues in annotation rules that remain inconsistent among Turkic UD treebanks. Tatar NMCTT Treebank is the first and the only UD treebank for Tatar, first released in UD v2.9. As of v2.12, it consists of 148 sentences and 2,280 tokens from online news articles. A characteristic unique to NMCTT treebank is that it explicitly annotates code-switched segments and their corresponding language code to mark Russian morphemes mixed in the text. This is important particularly because many other spoken Turkic languages commonly mix Russian due to language contact, and a similar annotation scheme can also be applied to them. With a special focus on the annotation of Tatar, the main issues discussed in this report are tokenization (e.g., locative adjectivizer -ki), part-of-speech (e.g., PART, converbs, -ki), morphological features (e.g., "Person" feature on default noun forms, "VerbForm" feature on converbs), and code-switching annotation.