Abstract

We present a Swedish-Turkish parallel treebank - a less processed and typologically dissimilar language pair - containing over 160,000 tokens in Swedish and 145,000 tokens in Turkish.

The treebank contains different annotation layers on the morphological and syntactic level using dependency structures.

The corpus is automatically created by reusing and adjusting existing tools for the automatic alignment and its visualization, and also partly manually corrected.

The treebank is currently used to teach students in Turkish, in linguistic research to study the two languages from a contrastive perspective, and in NLP applications.

Architecture