BOUN Treebank v2.11

Büşra Marşan, Furkan Akkurt, Onur Gungor, Tunga Güngör, Balkiz Ozturk, Suzan Uskudarli, Arzucan Özgür

The BOUN Treebank is created by the TABILAB, and supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under grant number 117E971 and by Boğaziçi University Research Fund under grant number 16909.

The BOUN Treebank includes a total of 9,761 manually annotated sentences from various topics including biographical texts, national newspapers, instructional texts, popular culture articles, and essays. The texts are taken from the Turkish National Corpus (TNC).

The dependency relations in the BOUN Treebank are manually annotated in the UD framework. The morphological features and UPOS information are first retrieved from the morphological parser of Sak et al. (2011) and converted to UD morphology automatically using our script. The morphological features, UPOS tags, XPOS tags, and lemma forms are then manually corrected in a systematic way.

v2.11 of the BOUN Treebank aims to tackle a set of issues that are caused by the discrepancies between UD annotation scheme and typology of Turkish language. Some of these issues are null morpheme copula, different functions of light verbs like ol-, rich derivational processes of Turkish, and segmentation of verbs with multiple TAM markers and/or copula.

Keywords: Universal Dependencies, Turkish treebank, Turkic languages, treebank, annotation, language resources