The more the merrier: a new dependency treebank for Turkish

This study reports on the largest and the most comprehensive treebank in Turkish with the efforts of 4 linguists and 5 NLP specialist. We annotated a subset of Turkish National Corpus. Our treebank introduces 9,761 authentic Turkish sentences from 5 different text types: essays, broadsheet national newspapers, instructional texts, popular culture articles, and biographical texts. In the annotation of the BOUN Treebank, we encoded the syntactic and the morphological informations using the Universal Dependencies (UD) framework which has been originated with the works of De Marneffe et al. (2014) and Nivre et al. (2016). The raw text acquired from the corpus is first translated into the UD format, and all the sentences are manually annotated over the course of nine months. We also followed the previous unification of annotation schemes for Turkish within the UD framework. We provide detailed guidelines and justification for our improvements and decisions within the annotation process. Moreover, we provide a new annotation tool specifically designed for agglutinative languages. All of our data, history of changes, tools regarding the manual annotation, and scripts for error checks are available online (link is hidden due to blind review). The annotated treebank can be used in various treebanks. In this study, we report improvements in the results of an state-of-the-art dependency parser in identifying heads and syntactic relations within sentences. We also report overall improvement when the parser is fed with all of the existing UD Treebanks including our new treebank. Inter-annotator agreement and parsing scores for different text types can be seen in Table 1 and Table 2. An example sentence from our treebank is presented in Figure 1, and a screenshot of tool can be seen in Figure 2.

References

De Marneffe, M. C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014, May). Universal Stanford dependencies: A cross-linguistic typology. In LREC (Vol. 14, pp. 4585-4592).

Nivre, J., De Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveria, N., Tsarfaty, R., Zeman, D. (2016, May). Universal dependencies v1: A multilingual treebank collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 1659-1666).

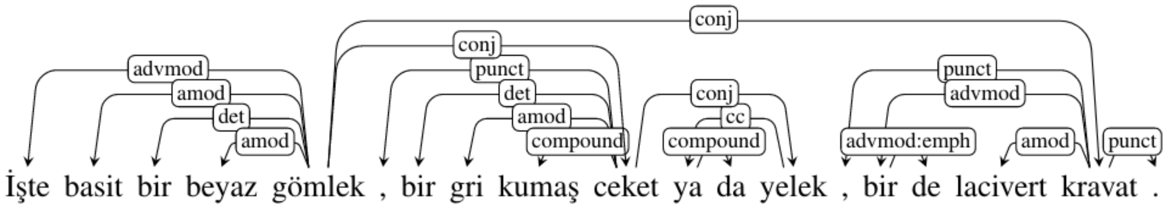
Aksan, Y., Aksan, M., Koltuksuz, A., Sezer, T., Mersinli, Ü., Demirhan, U. U., Yılmazer, H., Kurtoglu, Ö., Atasoy, G., Öz, S., Yıldız, İ. (2012, May). Construction of the Turkish National Corpus (TNC). In LREC (pp. 3223-3227).

Annotator Pair	κ_{Head}	κ_{Label}
1-2	0.82	0.83

Table 1: The Kappa measures of inter-annotator agreement with regards to head-dependent relation (κ_{Head}) and dependency tags (κ_{Label}).

Treebank	***Number of Sentences***	***UAS***	***LAS***
Essays	1,953	63.83	54.51
National Newspaper	1,898	~62	~53
Instructional Texts	1,976	73.95	65.14
Popular Culture Articles	1,962	74.18	65.89
Biographical Texts	1,972	~72	~64
New Treebank (Total)	9,761	~70	~60
Re-annotated IMST	5,635	75.49	65.53
Re-annotated PUD	1,000	78.70	70.01
All Treebanks	16,396	~79	~69

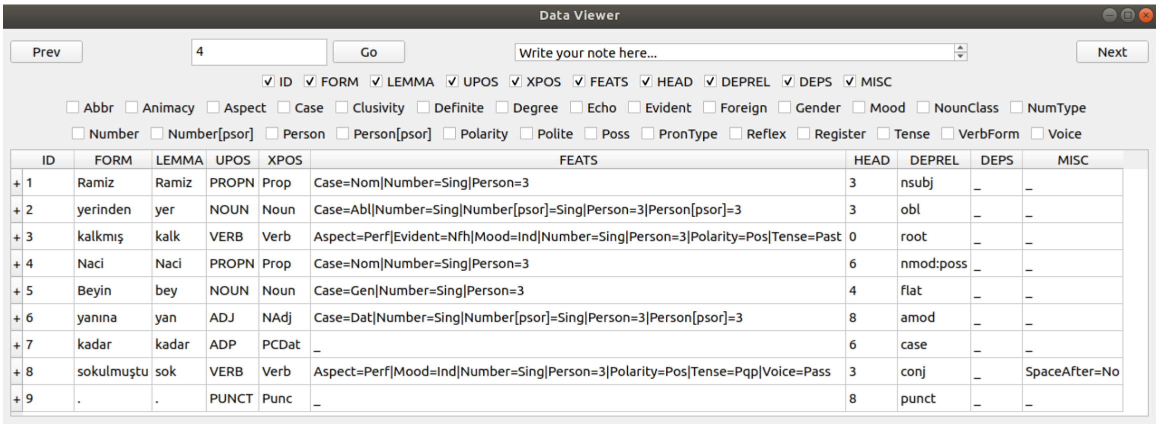
Table 2: UAS and LAS scores of the parser on each of the five sections of the our Treebank and the score for the entire treebank, as well as re-annotated IMST, PUD and the total of all Turkish UD treebanks. Scores with tilde symbol reports approximate scores.



İşte basit bir beyaz gömlek, bir gri kumaş ceket ya da yelek, bir de mavi kravat.
see basic a white shirt, a gray fabric jacket or FOC vest, a FOC blue tie.

'See, a simple white shirt, a gray blazer or a vest, and also a blue tie.'

Figure 1: An example sentence from our new treebank.



☑ ID ☑ FORM ☑ LEMMA ☑ UPOS ☑ XPOS ☑ FEATS ☑ HEAD ☑ DEPREL ☑ DEPS ☑ MISC

□ Abbr □ Animacy □ Aspect □ Case □ Clusivity □ Definite □ Degree □ Echo □ Evident □ Foreign □ Gender □ Mood □ NounClass □ NumType

□ Number □ Number[psor] □ Person □ Person[psor] □ Polarity □ Polite □ Poss □ PronType □ Reflex □ Register □ Tense □ VerbForm □ Voice

<table>
<tr><th>ID</th><th>FORM</th><th>LEMMA</th><th>UPOS</th><th>XPOS</th><th>FEATS</th><th>HEAD</th><th>DEPREL</th><th>DEPS</th><th>MISC</th></tr>
<tr><td>+ 1</td><td>Ramiz</td><td>Ramiz</td><td>PROPN</td><td>Prop</td><td>Case=Nom|Number=Sing|Person=3</td><td>3</td><td>nsubj</td><td>_</td><td>_</td></tr>
<tr><td>+ 2</td><td>yerinden</td><td>yer</td><td>NOUN</td><td>Noun</td><td>Case=Abl|Number=Sing|Number[psor]=Sing|Person=3|Person[psor]=3</td><td>3</td><td>obl</td><td>_</td><td>_</td></tr>
<tr><td>+ 3</td><td>kalkmış</td><td>kalk</td><td>VERB</td><td>Verb</td><td>Aspect=Perf|Evident=Nfh|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Past</td><td>0</td><td>root</td><td>_</td><td>_</td></tr>
<tr><td>+ 4</td><td>Naci</td><td>Naci</td><td>PROPN</td><td>Prop</td><td>Case=Nom|Number=Sing|Person=3</td><td>6</td><td>nmod:poss</td><td>_</td><td>_</td></tr>
<tr><td>+ 5</td><td>Beyin</td><td>bey</td><td>NOUN</td><td>Noun</td><td>Case=Gen|Number=Sing|Person=3</td><td>4</td><td>flat</td><td>_</td><td>_</td></tr>
<tr><td>+ 6</td><td>yanına</td><td>yan</td><td>ADJ</td><td>NAdj</td><td>Case=Dat|Number=Sing|Number[psor]=Sing|Person=3|Person[psor]=3</td><td>8</td><td>amod</td><td>_</td><td>_</td></tr>
<tr><td>+ 7</td><td>kadar</td><td>kadar</td><td>ADP</td><td>PCDat</td><td>_</td><td>6</td><td>case</td><td>_</td><td>_</td></tr>
<tr><td>+ 8</td><td>sokulmuştu</td><td>sok</td><td>VERB</td><td>Verb</td><td>Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Pqp|Voice=Pass</td><td>3</td><td>conj</td><td>_</td><td>SpaceAfter=No</td></tr>
<tr><td>+ 9</td><td>.</td><td>.</td><td>PUNCT</td><td>Punc</td><td>_</td><td>8</td><td>punct</td><td>_</td><td>_</td></tr>
</table>

Figure 2: A tabular view of our tool. Tree view is created under the tabular view.