

The more the merrier: a new dependency treebank for Turkish

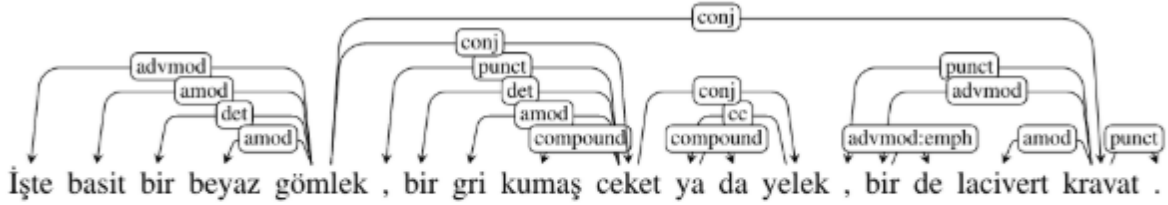
This study reports on the largest and the most comprehensive treebank in Turkish (BOUN Treebank) with the efforts of 4 linguists and 5 NLP specialist. We annotated a subset of Turkish National Corpus (Aksan et al., 2012). Our treebank introduces 9,761 authentic Turkish sentences from 5 different text 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 was first translated into the UD format, and all the sentences were 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 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 applications. In this study, we report improvements in the results of a state-of-the-art dependency parser in identifying heads and syntactic relations within sentences. We also report the overall improvement when the parser is fed with all 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 respectively. An example sentence from our treebank is presented in Figure 1 and a screenshot of the annotation tool can be seen in Figure 2.

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 and dependency tags.

Treebank	Number of Sentences	UAS	LAS
Essays	1,953	63.83	54.51
National Newspaper <input type="checkbox"/>	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
+1	Ramiz	Ramiz	PROFN	Prop	Case=Nom Number=Sing Person=3	3	nsubj	--	--
+2	yerinden	yer	NOUN	Noun	Case=Ab Number=Sing Number[psor]=Sing Person=3 Person[psor]=3	3	obl	--	--
+3	kalkmış	kalk	VERB	Verb	Aspect=Perf Evident=N Mood=Ind Number=Sing Person=3 Polarity=Pos Tense=Past	0	root	--	--
+4	Naci	Naci	PROFN	Prop	Case=Nom Number=Sing Person=3	6	nmod:poss	--	--
+5	Beyin	bey	NOUN	Noun	Case=Gen Number=Sing Person=3	4	flat	--	--
+6	yanna	yan	ADJ	NAdj	Case=Dat Number=Sing Number[psor]=Sing Person=3 Person[psor]=3	8	amod	--	--
+7	kadar	kadar	ADP	PCDat	--	6	case	--	--
+8	sakulmuştu	sek	VERB	Verb	Aspect=Perf Mood=Ind Number=Sing Person=3 Polarity=Pos Tense=Prog Voice=Pass	3	conj	--	SpaceAfter=No
+9	.	.	PUNCT	Punc	--	8	punct	--	--

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

Keywords: Turkish, dependency, treebank, corpus, manual annotation

Kaynakça/References (APA format)

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

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., ... & Tsarfaty, R. (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).