Turkish Treebanking: Unifying and Constructing Efforts

Utku Türk‡, Furkan Atmaca‡, Şaziye Betül Özateş*, Abdullatif Köksal*, Balkız Öztürk‡, Tunga Güngör*, Arzucan Özgür*

‡Department of Linguistics, *Department of Computer Engineering,
Boğaziçi University, Bebek, 34342 İstanbul, Turkey
{utku.turk,furkan.atmaca,saziye.bilgin,abdullatif.koksal,balkiz.ozturk,gungort,arzucan.ozgur}@boun.edu.tr

Text Analytics and BioInformatics Laboratory

Turkish Dependency Treebanks

Treebank	Size(sent.)	Status	Re-annotation	Name
IMST-UD [6]	5635	Finished	Done	BIMST
PUD [8]	1000	Finished	Done	BPUD
UD_Turkish-GB [2]	2802	Finished	Excluded	-
TNC-UD	10,000	Continued	-	-

We present;

- the first annotation of the Turkish National Corpus Universal Dependencies (TNC-UD) Treebank
- the re-annotation of the Turkish PUD Treebank
- a custom annotation software with advanced filtering and morphological editing options.

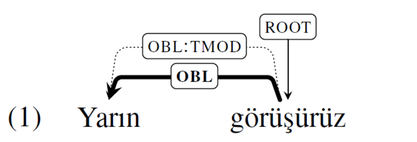
Re-annotating Turkish PUD Treebank

- Re-annotation of Turkish PUD Treebank was done in conjunction with the re-annotation of the IMST-UD Treebank. The re-annoated version of IMST-UD, the BIMST Treebank[7] can be found with its guidelines at `https://github.com/boun-tabi/UD_Turkish-BIMST`.

Consistency related changes:

We simplified the language specific syntactic relation tags that are used in Turkish PUD Treebank, but not in IMST-UD.

- *obl:tmod* to *obl*, *acl:relcl* to *acl*, *det:predet* to *det*, *flat:name* to *flat*



(1) Yarın görüşürüz

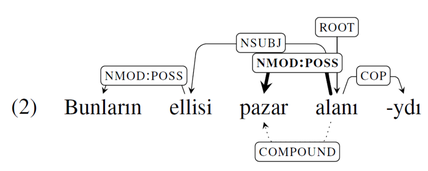
Yarın gör-üş-ür-üz
tomorrow see-RCP-AOR-3SG.PL

'See you tomorrow'

The dotted lines show the relations used in the previous treebank, the bold ones indicate the re-annotated ones in the updated version, and the fine lines represent unaltered dependencies.

Linguistically driven changes:

- Compound to nmod:poss relations:

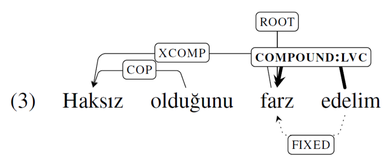


(2) Bunların ellisi pazar alanı -ydı

Bun-lar-ın elli-si pazar alan-ı-ydı
this-PL fifty-POSS market place-POSS-COP

'50 of these were marketplaces'

- Fixed to compound:lvc relations:

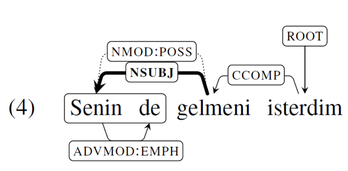


(3) Haksız olduğunu farz edelim

Haksız ol-duğ-un-u farz ed-elim
wrongful be-NMLZ-POSS-ACC presume do-OPT

'Let's just say he's wrong.'

- Nmod:poss to nsubj relations:



(4) Senin de gelmeni isterdim

Sen-in de gel-me-ni
you-GEN too come-NMLZ-POSS
iste-r-di-m.
want-AOR-PST-1SG

'I would have wanted you come, as well.'

The full history of the re-annotation of the Turkish PUD Treebank and its final version, named as Turkish BPUD Treebank are available at `https://github.com/boun-tabi/UD_TURKISH-BPUD`.

The TNC-UD Treebank

Manually annotated sentences from TNC Corpus [1].

- Randomly selected 10 thousand sentences from essays, broadsheet national newspapers, instructional texts, popular culture articles, and biographical texts registers.
- Morphological analyses of the sentences were created by the Turku Neural Parser Pipeline [5].
- Current version contains 500 annotated sentences, annotation of additional 9,500 sentences is in progress.

Annotating the Treebanks

- A team of three linguists and four computer scientists
- Created an annotation guideline for Turkish based on [3].
  - every detail is discussed by the entire group.
  - publicly available at `https://github.com/boun-tabi/UD_TURKISH-BPUD`.
- Inter-annotator agreement is 99% for finding correct heads and 88% for finding correct dependency labels.

Experiments

We used a state-of-the-art neural parser [4] in the experiments.

- 5-fold cross-validation is used: each sub-part includes 200 sentences for PUD and BPUD treebanks.
- In the evaluation, the unlabeled attachment score (UAS) and the labeled attachment score (LAS) metrics are used.

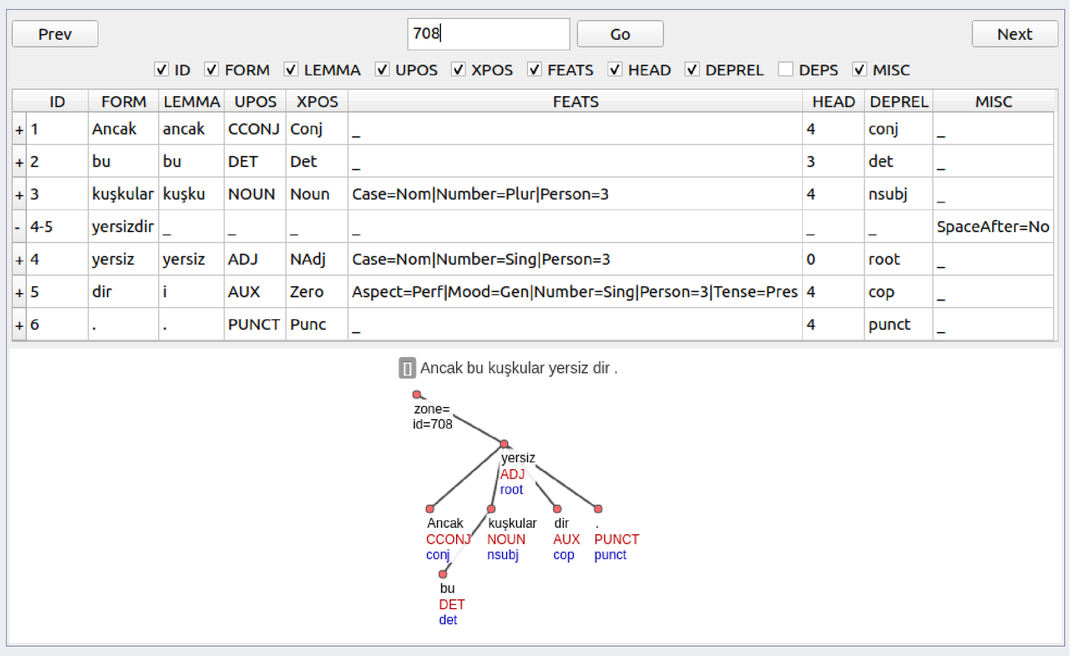
Treebank	UAS	LAS
PUD	**79.83**	**74.31**
BPUD	78.70	70.01
BPUD & TNC-UD	**79.79**	**71.22**
IMST-UD & PUD	**82.41**	**77.47**
BIMST & PUD	81.77	73.68

The differences in the attachment scores might result from the annotation scheme adopted in this study.

- Our main aim is to ensure consistent and linguistically correct annotations that follow the UD guidelines.
- We enhanced the annotations of the treebanks that have previously rough and incorrect annotations.
- These more accurate annotations of the treebanks will lead to better and more consistent parsing accuracies when more annotated data is available.

BoAT - BOUN Annotation Tool for Dependency Parsing

The BoAT annotation tool is an open-source desktop application written in Python3 with PyQt5 library.



- A comfortable and intuitive environment for annotators.
- The ability to declutter the working environment.
- The automatic validation process.
- An easy way to edit multiword expressions.

These abilities of our tool make it one of the first tools that is shaped according to the needs of the Turkish language. It is available at `https://github.com/boun-tabi/BoAT`.

Acknowledgements

This work was supported by the Scientific and Technological Research Council of Turkey(TÜBITAK) under grant number 117E971 and as a graduate scholarship.

References

[1] Y. Aksan, M. Aksan, A. Koltuksuz, T. Sezer, Ü. Mersinli, U. U. Demirhan, H. Yılmazer, G. Atasoy, S. Öz, İ. Yıldız, and Ö. Kurtoğlu. Construction of the Turkish National Corpus (TNC). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3223–3227, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[2] C. Cöltekin. A grammar-book treebank of turkish. In *Proceedings of the 14th workshop on Treebanks and Linguistic Theories (TLT 14)*, pages 35–49, 2015.

[3] M.-C. de Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, and C. D. Manning. Universal Stanford Dependencies: A cross-linguistic typology. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 4585–4592, 2014.

[4] T. Dozat, P. Qi, and C. D. Manning. Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30, 2017.

[5] J. Kanerva, F. Ginter, N. Miekka, A. Leino, and T. Salakoski. Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 133–142, 2018.

[6] U. Sulubacak, M. Gökırmak, and F. M. Tyers. Universal Dependencies for Turkish. *Proceedings of the 26th International Conference on Computational Linguistics (COLING-16)*, pages 3444–3454, 2016.

[7] U. Türk, F. Atmaca, B. Özateş, B. Öztürk, T. Güngör, and A. Özgür. Improving the Annotations in the Turkish Universal Dependency Treebank. In *Universal Dependencies Workshop (UDW 2017) at SyntaxFest 2019.*

[8] D. Zeman, F. Ginter, J. Hajič, J. Nivre, M. Popel, M. Straka, and et al. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–20. Association for Computational Linguistics, 2017.