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

Abstract

In this paper, we present the re-annotation of the Turkish PUD Treebank and the first annotation of the Turkish National Corpus Universal Dependency (henceforth TNC-UD) Treebank as part of our efforts for unifying and extending the Turkish universal dependency treebanks. In accordance with the Universal Dependencies' guidelines and the necessities of Turkish grammar, both treebanks, the Turkish PUD Treebank and TNC-UD, were revised with regards to their syntactic relations. The TNC-UD is planned to have 10,000 sentences. In this paper, we present the first 500 sentences along with the re-annotation of the PUD Treebank. Moreover, this paper also offers the parsing results of a graph-based neural parser on the previous and re-annotated PUD, as well as the TNC-UD. In light of the comparisons, even though we observe a slight decrease in the attachment scores of the Turkish PUD treebank, we demonstrate that the annotation of the TNC-UD improves the parsing accuracy of Turkish. In addition to the treebanks, we have also constructed a custom annotation software with advanced filtering and morphological editing options. Both of the treebanks, including a full edit-history and the annotation guidelines, as well as the custom software are publicly available online under an open license.

1 Introduction

The Universal Dependency (UD) project has proven itself to be an indispensable part of the natural language processing (NLP) framework. The treebanks built within the scope of the project constitute a great portion of the contribution made by the UD Project to NLP applications. However, within the UD Project, there is a significant mismatch regarding the volume of the treebanks available for each language. Turkish is one of the under-resourced languages; even though previous treebanks (Sulubacak et al., 2016a) do exist together with works on Turkish morphology (Çöltekin, 2016, 2015), the limited number of Turkish resources poses a challenge for those who wish to conduct NLP studies.

The main contribution of this paper is making up for the scarcity of NLP resources in Turkish by annotating a new corpus that has not been introduced to the UD project before, namely the TNC (Aksan et al., 2012). The current version of the annotated treebank only contains 500 sentences; however, we are currently working to an additional 9,500 sentences to the corpus. The syntactic relations of the sentences in the treebank were manually annotated following the Stanford Dependency (SD) scheme (de Marneffe et al., 2014) as well as the UD guidelines. Moreover, the morphological analyses of the sentences were automatically created by the Turku Neural Parser Pipeline (Kanerva et al., 2018) trained on the re-annotated version of the Turkish IMST-UD Treebank that we are currently working on.

As a second contribution, we manually re-annotated the Turkish PUD treebank for consistency in the annotation. As we do not fully agree with the annotation scheme of previous Turkish treebanks, we had incorporated a more strict view of the SD scheme and tried to balance the six directives of Manning's Law (Nivre et al., 2017). Our objective is to unify the annotation schemes and the level of granularity in terms of linguistic depth within the Turkish treebanks in the UD Project. The linguistic decisions and departures from the previous work related to Turkish tree-

Proceedings of the 13th Linguistic Annotation Workshop, pages 166–177
Florence, Italy, August 1, 2019.

banks will also be exemplified in this paper.

As a third contribution, we present an open source desktop application for the annotation process. Our proposed annotation tool integrates a tabular view, a hierarchical tree structure which can also be read in a linear fashion, and advanced morphological editing and filtering features. The tabular aspect of the annotation tool enables a keyboard-driven process for annotators; thus, helping with speed and ergonomy related problems by getting rid of the excessive use of the mouse. The linearity and the hierarchical view are inspired by the CoNLL-U Viewer, which helps linguists in visualizing the data.

2 Related Work

Within the last decade of the 20th century, treebanks and annotated corpora started to hold an extremely important place for NLP tools, applications, and scientific research within the framework of NLP. Even though creating such corpora that are structurally consistent and big enough to help NLP processes was incredibly tedious and time-consuming, it was believed to be worth pursuing by many.

Emulating the first efforts to create an annotated treebank from a corpus in English and in other languages (Marcus et al., 1993; Böhmová et al., 2003; Taylor et al., 2003), Oflazer et al. (2003) and Atalay et al. (2003) introduced the first Turkish treebank, the METU-Sabancı Treebank (MST), consisting of 5,635 sentences. A majority of the sentences in this treebank were drawn from either newspaper articles or novels, making up 42% and 13% of the corpus, respectively. Even though it may seem that the register of the treebank is overwhelmingly newspaper oriented, no other Turkish treebank matched it in size.

The other important aspect of MST was the fact that it became the originating point for the first Turkish UD treebank, IMST-UD. Firstly, Sulubacak et al. (2016b) revisited the syntactic and morphological decisions made in MST, and re-annotated the treebank from the ground up. Unlike Atalay et al. (2003), Sulubacak et al. (2016b) provided the necessary information regarding the annotation process, such as the number of annotators, their background, and the decision making mechanism for the ITU-METU-Sabancı Treebank (IMST). However, their work still lacked inter-annotator agreement scores and a description of the process behind finding solutions to disagreements, which makes up one of the most important aspects of building an annotated treebank.

After the creation of IMST, Sulubacak et al. (2016a) automatically converted it into the first Turkish treebank in a UD release, resulting in unparalleled success with respect to the attachment scores. They also provide a thorough description of mappings and the automation process. However, the treebank was not post-edited by a human. Until the very recent edits, the treebank had very problematic consistency issues as well as a faulty representation, i.e. punctuations as roots, reversed head-dependent relations etc., caused by the automated nature of the conversion into the UD framework. Even though four different updates were made to the IMST-UD and most of the problems are now resolved, there are still vital divergences from the SD scheme and UD guidelines, such as a non-satisfactory distinction between core and non-core dependencies, the inner structure of embedded clauses, and multiword expressions that include the morphologically ambiguous *-ki* marker (Çöltekin, 2016).

Apart from this line of work, Turkish also has two other annotated treebanks within the UD framework: the Grammar-Book Treebank (Çöltekin, 2015) and the Turkish PUD Treebank[1]. Even though it offers a grand resource containing 2,803 sentences, we excluded the Grammar-Book Treebank in our research for consistency related reasons. The sentences in the Grammar-Book Treebank are unnatural with regards to grammaticality. In other words, the sentences in the treebank are either perfectly good sentences that are engineered to be grammatical, short, and concise or they are fragments of sentences that cannot stand alone and are unlikely to be uttered in isolation.

As for the Turkish PUD Treebank, it was published as a part of the CoNLL 2017 Shared Task on Multilingual Parsing from Raw Text to Universal Dependencies (Zeman et al., 2017). It consists of 1,000 sentences that were parallel annotated for 18 languages, 14 of them used in the shared task. One of the biggest contributions of this treebank is that it allows researchers in the NLP framework to reach a solid common ground in terms of items. Moreover, it allowed researchers to get rid of problems such as hidden semantic or individ-

[1] universaldependencies.org/treebanks/tr_pud/

ual sentence related confounds. However, unlike languages like English which are manually annotated in native UD, the Turkish PUD Treebank was not annotated manually in native UD style. The process is fairly similar to the creation of the IMST-UD, which involves non-UD style annotation followed by automatic conversion into the UD style. This automatic conversion includes Universal POS, features, and relations. Moreover, much like the IMST-UD, the Turkish PUD Treebank also lacks crucial information like annotator information, inter-annotator agreement, annotation process, and any post-editing process.

In contrast to most of the other annotated treebanks in the UD Project, Turkish treebanks yield an inconsistent picture with regards to their underlying annotations. Furthermore, they lack explanatory information about the annotation process and none were annotated in the native UD style. Thus, it is almost impossible to consider the Turkish treebanks in the UD Project as one unified and structured treebank.

Considering the development of Turkish treebanks in the UD Project, the next most logical step was to first investigate the automatic conversion process and re-annotate the treebanks. We are currently working on re-annotating IMST-UD. The main problems which we yet encountered in the process of re-annotation of IMST-UD can be grouped into three important group: the analyses of embedded clauses, the discussion of core and non-core arguments, and the newly introduced dependency types. Due to the nature of automatic conversion, IMST-UD lacked the necessary linguistic depth with regards to embedded structure. Instead of a hierarchical representation of inner argument and event structure, they were represented as simple nominal phrases. This was due to the nature of the nominalization phenomenon that is present in almost all Turkic and Altaic languages. Moreover, the IMST-UD was criticized for not differentiating between core elements that are non-canonically case-marked and adjuncts using the same case markers. Turkish makes use of cases except the accusative case to mark the core dependents of the predicate. Different than `obl`, when these dependents are left out of the sentence, sentences either gain a totally different meaning or become ungrammatical to native speakers of Turkish. Lastly, we included eight new syntactic relations that are used in UD v2.0, but not used in the IMST-UD. The details of these issues will be explained thoroughly in future work.

Due to the lack of a coherent picture in Turkish treebanks, Turkish NLP tools and applications have remained scarce and stagnant. TRmorph, ITU Treebank Annotation Tool, and the annotation tool of Atalay et al. (2003) are some of the few available tools in this field (Çöltekin, 2014; Eryiğit, 2007; Çöltekin, 2010).

With these reasons in mind, we decided to unify the approach towards the Turkish treebanks within the UD framework. With this initiative, we aimed to create a more consistent picture of Turkish for NLP tools and applications and enhance the use of Turkish in various NLP tasks. As aforementioned, we re-annotated the Turkish PUD Treebank and introduced the first steps to the creation of a new treebank: the TNC-UD.

3 Re-annotating Turkish PUD Treebank

Even though the Turkish PUD Treebank offers a much cleaner picture than the IMST-UD, it is not without its erroneous annotation. However, before addressing the errors, we will discuss the changes that we implemented for the sake of consistency in the two Turkish treebanks.

The consistency related changes mostly include the simplification of the language specific syntactic relation tags that are used in the Turkish PUD Treebank, but not in IMST-UD. We believe that in cases like Example 1, the syntactic relation of `obl` is a sufficient annotation in terms of linguistic adequacy. Such cases include changes from `obl:tmod`, `acl:relcl`, `det:predet`, `flat:name` syntactic relations to `obl`, `acl`, `det`, `flat`, respectively.

(1) Yarın görüşürüz

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

'See you tomorrow'

Having tackled the consistency related issues, we can turn our focus to the linguistically driven

[1]In all dependency trees in this paper, the dotted lines show the syntactic relations used in the previous treebank, the bold ones indicate the re-annotated ones in the updated treebank, and the fine lines represent unaltered dependencies.

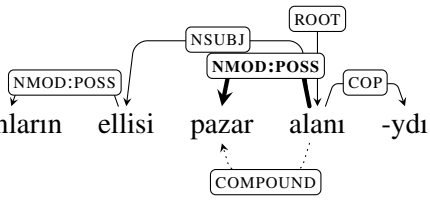
changes. Table 1 shows the most frequently applied changes, excluding the changes made for reasons solely driven out of consistency.

Turkish PUD Treebank	Boğaziçi PUD Treebank	Number of Alterations
COMPOUND	NMOD:POSS	1331
NMOD:POSS	NSUBJ	192
FIXED	COMPOUND:LVC	168

Table 1: The number of alterations that we made for the most frequent changes.

As is evident in Table 1, the change from the syntactic relation `compound` to `nmod:poss` is overwhelmingly high. It makes up 28% of the total changes. Apart from the most changed three syntactic relations the rest was not even close to these changes in number.

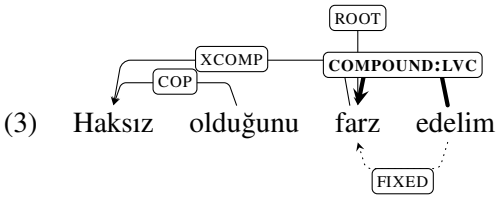
It is no surprise that compounds are in the spotlight in these changes. Compounds have always been a controversial topic in Turkish (Hayasi, 1996; Swift, 1963; Göksel, 2009; Göksel and Haznedar, 2007; Göksel and Kerslake, 2005; Öztürk and Erguvanlı-Taylan, 2016). Within the UD guidelines and SD scheme, compounds are treated as head-level (X^0) constructions, which is different than Noun+Noun (NN) constructions that have syntactic reflex in the phrasal level and from compounds that are lexicalized with time. However, the Turkish PUD Treebank does not distinguish between these constructions. As seen in Example 2, the existence of a syntactic reflex, possessive marker, on *alan-ı-ydı* indicates the phrasal level of construction. Since possessive marker in Turkish introduces a transitivity relation, we can conclude that apart from lexicalized NN-*(s)I(n)* constructions, are not head-level constructions. This is why, in the re-annotation process, we have carried out a great number of alterations from the syntactic relation `compound` to `nmod:poss`.



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

'50 of these were marketplaces'

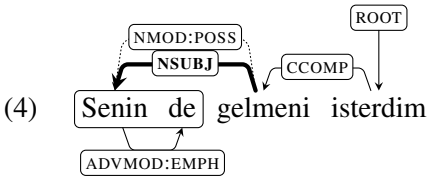
Following the discussion of compounds, the light verb constructions were also problematic in the Turkish PUD Treebank as seen in Example 3. They were annotated as `fixed`, instead of `compound:lvc` which is highly used in IMST-UD as well as in treebanks of other languages like Persian and Armenian (Seraji et al., 2016; Yavrumyan et al., 2017). The analysis follows from the fact that even though light verbs are grammaticalized expressions, they do still have an internal structure, which separates them from being `fixed` according to the UD guidelines.



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

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

The second most frequent alteration overlaps with the issues that have been addressed in the re-annotation of the IMST-UD. The sentence given in Example 4 is an example of the lack of the inner structure of an embedded sentence. Even though it is marked with genitive case, which is the canonical way of marking `nmod:poss` in Turkish, *senin de* is not just a possessive nominal modifier; instead, it is the subject of the embedded clause.



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.'

4 Turkish National Corpus UD Treebank

In the current version of the planned treebank, the sentences are drawn from the Turkish National Corpus (TNC) (Aksan et al., 2012). The reason why we selected our sentences from TNC is based on our preference for freely available corpora. TNC is free to use for research purposes and it includes 5 million words of written texts across a variety of genres.

Even though the original TNC corpus has 22 main registers, we only included sentences from 5 different registers: essays, broadsheet national newspapers, instructional texts, popular culture articles, and biographical texts. Each register contributes to the total treebank with 2,000 sentences, which corresponds to 25% of the treebank. Sentences were drawn randomly from these registers with the help of those who regulate the corpus.

The motivation for the selection of these text types was based on the linguistic variety and the integrity of the texts with regards to grammaticality so that it does not hinder the annotation process. The selection of registers includes sentences with an evenly distributed variety of length (from essays to instructional texts), formality (from newspapers to popular culture articles), and literary quality (from biographical texts to newspapers).

We also obtained 5,000 sentences from non-academic texts about natural sciences, humanities, social sciences, medicine, and engineering. These sentences will only be used in case of exclusions from the original 10,000 sentences. In case of such an exclusion, sentences from the non-academic text pool will be randomly selected and annotated in order to reach the target of 10,000 sentences.

5 Annotating the Treebanks

For the re-annotation of the Turkish PUD Treebank and the annotation of the TNC-UD, we used a team of two annotators who are linguists and have comprehensive knowledge of Turkish grammar and general linguistics, as well as grammatical theories. Supporting the team of annotators, we have a senior linguist who leads the discussion whenever there is a disagreement between the two annotators. In addition to the three linguists, a team of four computer scientists with considerable experience in NLP research monitored the process of manual annotation.

As a first step, we created a guideline of annotation in the native UD style and SD scheme. We used the already existing guidelines as a basis (de Marneffe et al., 2014) and focused on optimizing them, especially the guidelines that were created for Turkish. The guidelines were created after every detail was discussed by the entire group of three linguists and four computer scientists. The guidelines were then exemplified with possible sentences. These guidelines are made available together with the other relevant data, corpus, and the software.

Due to time and resource restrictions, we were unable to employ full double annotation. Instead, after each annotator completed his/her own part, the sentences were run through an adjudication process within the group of linguists. When a disagreement occurred, the team discussed it thoroughly before applying the last judgment consistently for all the similar examples. Double annotation was performed for a set of 300 randomly selected sentences. Table 2 shows the kappa measures of inter-annotator agreement for finding the correct heads (κ_{Head}) and the correct dependency label of the syntactic relations (κ_{Label}).

Annotator Pair	κ_{Head}	κ_{Label}
1-2	0.9966	0.8873

Table 2: The Kappa measures of inter-annotator agreement with regards to head-dependent relation and dependency tags.

6 Released Data and Software

With the release of the treebank, we also release the full history of the annotation of TNC-UD, as well as the full history of the re-annotation of the Turkish PUD Treebank. Furthermore, we plan to provide statistical figures about the changes we have employed. We believe that the full transparency and the full replicability of the results are extremely important.

As well as the data and the history of change, the release of the treebank also includes our improvements on the UD Guidelines for Turkish. These guidelines include the necessary explanations and sentences accompanied with theoretical discussion. Being able to trace back our decisions will enable us and other researchers to accommodate according to the new findings in both linguistics and NLP fields in the future.

Finally, we release a desktop annotation tool that is designed for linguists with the aim of advanced morphological editing, ease of use, and decluttering the working environment. Our annotation tool is an open-source desktop application written in Python3 with PyQt5 library. The main objective of the tool is to create a comfortable, fast, and intuitive environment for annotators. As shown in Figure 1, its tabular view enables annotators wander freely only using their

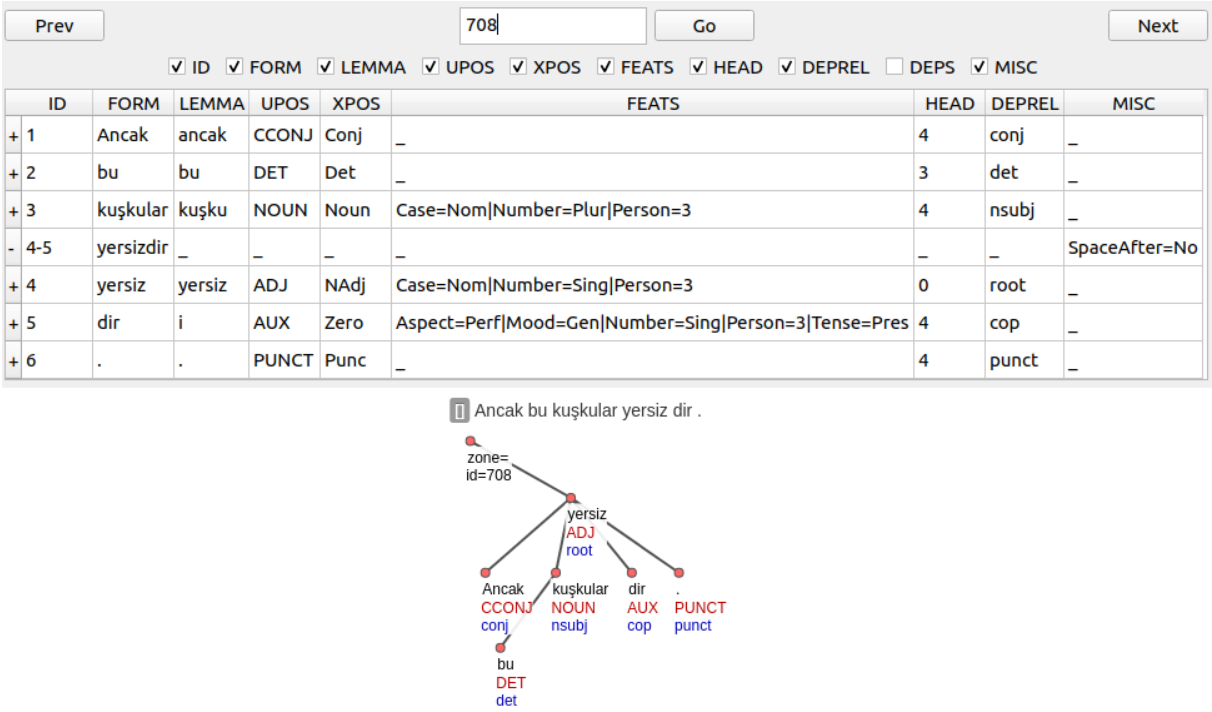


Figure 1: A screenshot of our annotation tool that integrates a tabular view with a hierarchical, but linearly readable tree. The plus and minus symbols on the left enable annotators to easily edit multiword expressions.

keyboard, which eliminates the hustle of using a mouse and the possible wrist injuries in the long hours of manual annotation.

Another important aspect of our tool is its ability to declutter the working environment. Annotators can change the information that is visible at a time with the check boxes above. It also facilitates the validation process since it checks the validity of the trees at every click of the Next and Prev buttons. If an erroneous annotation is detected, such as having two roots, one node having two parents, typos inside the tabs, and the like, it immediately gives an error and informs the annotator about the error.

Besides these, the main objective behind our tool is offering an easy way to edit multiword expressions in agglutinative languages like Turkish. In an automated conversion process, languages like Turkish may face a large number of erroneous tags with respect to multiword expressions. Editing those tags is extremely tedious since there is no way of keeping up with the dependencies and their heads. Our annotation tool enables annotators to easily split a word into two and also easily join them by pressing the plus and minus buttons. Upon such edits, every dependency relation and their ID's are automatically updated. Thus, these abilities of our tool make it one of the first tools that is shaped according to the needs of the Turkish language.

Lastly, we have ported the CoNLL-U viewer to our annotation tool by changing the related methods in the UDAPI library (Popel et al., 2017). Its hierarchical, yet linearly readable approach is intuitive to many linguists who work in the annotation processes.

7 Experiments

To see the effect of re-annotation on the parsing accuracy, we trained a state-of-the-art graph-based neural parser (Dozat et al., 2017) on the previous and re-annotated versions of the PUD and TNC-UD treebanks. Due to the insufficient amount of data, we use the 5-fold cross-validation technique on the Turkish PUD treebank where each sub-part includes 200 sentences. So the training data size is 600 sentences, and the sizes of the development and test sets is 200 sentences in each fold. To evaluate the TNC-UD Treebank, we trained a model where the TNC-UD Treebank is used as an additional training data for the re-annotated version of the PUD Treebank and then the trained model is evaluated on the test set of the PUD Treebank. We again use the 5-fold cross-validation technique

to evaluate this setting. Both projective and non-projective dependencies are included in the training and test phases.

In the evaluation of the dependency parser, we used the word-based unlabeled attachment score (UAS) metric, which is measured as the percentage of words that are attached to the correct head, and the labeled attachment score (LAS) metric, which is defined as the percentage of words that are attached to the correct head with the correct dependency type.

In all of the tables that show the results of the experiments performed, the attachment scores of the parser on both the previous version and the re-annotated version of the treebanks are given. Although comparing these scores is not a correct approach, since the test data sets that the models are evaluated on are annotated differently, observing the parsing accuracies of the previous and the re-annotated versions of the treebanks together gives a better idea to understand the current state of the parsing success of Turkish.

Table 3 shows the attachment scores of the parser on the previous and re-annotated versions of the Turkish PUD Treebank test data set. The re-annotated version of the Turkish PUD Treebank is named as BPUD.

Treebank	UAS	LAS
PUD	**79.83**	**74.31**
BPUD	78.70	70.01

Table 3: UAS and LAS scores of the parser on the previous and re-annotated versions of the Turkish PUD Treebank test data set when the parser is trained only with the training data set of the Turkish PUD Treebank.

From the results, we observe a decrease in the parsing accuracy in terms of the attachment scores. Although the decline in the UAS score is not large, the difference between the LAS scores of the two versions is four percent.

In order to understand whether these results are because of the insufficient amount of training data, we performed additional experiments by including the training set of the corresponding version (i.e., the previous version and the re-annotated version) of the Turkish IMST-UD Treebank to the training data of the PUD Treebank using the 5-fold cross-validation technique. In this setting, the training data set consists of 600 sentence PUD training set and 3685 sentence IMST-UD training set. The development set includes 200 sentence PUD development set and 975 sentence IMST-UD development set in each fold. The test set remains the same as in the previous experiment.

Table 4 depicts the UAS and LAS scores of the parser when both IMST-UD and PUD are included in the training phase.

Treebank	UAS	LAS
Previous version of IMST-UD & PUD	**82.41**	**77.47**
Updated version IMST-UD & PUD	81.77	73.68

Table 4: UAS and LAS scores of the parser on the previous and re-annotated versions of the Turkish PUD Treebank test data set when the parser is trained on the training data sets of the Turkish PUD Treebank and the IMST-UD Treebank.

We see that when we increase the size of the training data, the gap between the attachment scores gets smaller between the previous and re-annotated versions of the Turkish PUD Treebank.

The differences in the attachment scores of the previous and the re-annotated versions might result from the annotation scheme adopted in this study. In the re-annotation process, our main aim is to ensure consistent and linguistically correct annotations that follow the UD guidelines. By doing this, we enhanced and elaborated the annotations of the treebanks that have previously rough and incorrect annotations. So, when there is not sufficient amount of training data, the task of learning the syntactic relations between the words of a sentence is harder on the re-annotated versions of the treebanks. The experimental results suggest that, these more accurate annotations of the treebanks will lead to better and more consistent parsing accuracies when more annotated data is available.

We also made an experiment to see the impact of the TNC-UD Treebank on the parsing accuracy of the parser. Table 5 shows the attachment scores when the parser is trained on the PUD and TNC-UD treebanks.

Treebank	UAS	LAS
BPUD & TNC-UD	**79.79**	**71.22**
BPUD	78.70	70.01

Table 5: UAS and LAS scores of the parser on the re-annotated version of the Turkish PUD Treebank test data set when the parser is trained with the training data set of the Turkish PUD Treebank and the TNC-UD Treebank.

Even though the current version of the TNC-

UD Treebank includes only 500 annotated sentences, the parsing performance of the parser has increased more than 1 point in terms of the attachment scores.

The experiment results suggest that the final version of the TNC-UD Treebank which will consist of 10,000 annotated sentences together with the other linguistically corrected Turkish treebanks will greatly improve the syntactic parsing of Turkish texts.

8 Conclusion and Future Work

In this work, we have presented the re-annotation of the Turkish PUD Treebank and the first steps of annotating the TNC-UD Treebank, a new freely available treebank for Turkish. We believe that we have unified the annotation style of the Turkish treebanks in the UD framework. Moreover, we plan to annotate a total of 10,000 sentences in the native UD style, following the SD scheme (de Marneffe et al., 2014). The TNC-UD Treebank consists of four sections, with texts from different registers: essays, broadsheet national newspapers, instructional texts, popular culture articles, and biographical texts.

In the TNC-UD Treebank, morphological analyses has been provided with a deep learning-based parser pipeline (Kanerva et al., 2018) trained on the re-annotated version of the Turkish IMST-UD Treebank. In the syntactic analyses, we have used a team of two linguists for manual annotation. The inter-annotator agreement was 99% and 88% for finding correct heads and correct dependency label of the syntactic relations, respectively. This level of high agreement shows that both annotators followed the pre-prepared guidelines and examples with SD scheme strictly.

The annotated treebanks, the detailed history of changes made in the annotation process, and our new guidelines are available at `https://github.com/boun-tabi/UD_TURKISH-BPUD`. Moreover, our desktop annotation tool is available at `https://github.com/boun-tabi/BoAT`

Our current goal is to complete the annotation of the TNC-UD Treebank. We believe that 10,000 sentences manually annotated in the native UD style would enable NLP applications even more and help researchers to create a more robust environment for statistical learning.

One other future goal of this work is to enhance the annotation of the TNC-UD Treebank. Such annotation could include human-validated morphological analyses, prosodic information of the sentence, and detailed semantic analysis.

Acknowledgments

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

References

Yeşim Aksan, Mustafa Aksan, Ahmet Koltuksuz, Taner Sezer, Ümit Mersinli, Umut Ufuk Demirhan, Hakan Yılmazer, Gülsüm Atasoy, Seda Öz, İpek Yıldız, and Özlem Kurtoğlu. 2012. 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. European Language Resources Association (ELRA).

Nart Bedin Atalay, Kemal Oflazer, and Bilge Say. 2003. The annotation process in the Turkish treebank. In *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003*.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The Prague dependency treebank. In *Treebanks*, pages 103–127. Springer.

Çağrı Çöltekin. 2010. A freely available morphological analyzer for Turkish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 820–827.

Çağrı Çöltekin. 2014. A set of open source tools for Turkish natural language processing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1079–1086. European Language Resources Association (ELRA).

Çağrı Çöltekin. 2015. A grammar-book treebank of Turkish. In *Proceedings of the 14th workshop on Treebanks and Linguistic Theories (TLT 14)*, pages 35–49.

Çağrı Çöltekin. 2016. (When) do we need inflectional groups? In *Proceedings of The First International Conference on Turkic Computational Linguistics*.

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. 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.

Gülşen Eryiğit. 2007. ITU Treebank Annotation Tool. In *Proceedings of the ACL workshop on Linguistic Annotation (LAW 2007)*, Prague.

Aslı Göksel. 2009. Compounds in Turkish. *Lingue e linguaggio*, 8(2):213–236.

Aslı Göksel and Belma Haznedar. 2007. Remarks on compounding in Turkish. *MorboComp Project, University of Bologna*.

Aslı Göksel and Celia Kerslake. 2005. *Turkish: A Comprehensive Grammar*. Comprehensive grammars. Routledge.

Tooru Hayasi. 1996. The dual status of possessive compounds in modern Turkish. *Symbolae Turcologicae. Studies in honor of Lars Johanson on the occasion of his sixtieth birthday*, 6:119–29.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. 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.

Mitchell Marcus, Beatrice Santorini, and Mary Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, 19:330–331.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 4585–4592.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis M. Tyers. 2017. Tutorial on Universal Dependencies. Presented at European Chapter of the Association for Computational Linguistics, Valencia [Accessed: 2019 04 08].

Kemal Oflazer, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür. 2003. Building a Turkish Treebank. In *Treebanks, Building and Using Parsed Corpora*, pages 261–277.

Balkız Öztürk and Eser Erguvanlı-Taylan. 2016. Possessive constructions in Turkish. *Lingua*, 182:88–108.

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden. Association for Computational Linguistics.

Mojgan Seraji, Filip Ginter, and Joakim Nivre. 2016. Universal Dependencies for Persian. In *LREC*.

Umut Sulubacak, Memduh Gökırmak, and Francis M. Tyers. 2016a. Universal Dependencies for Turkish. *Proceedings of the 26th International Conference on Computational Linguistics (COLING-16)*, pages 3444–3454.

Umut Sulubacak, Tugba Pamay, and Gülşen Eryiğit. 2016b. IMST: A Revisited Turkish Dependency Treebank. In *In Proceedings of 1st International Conference on Turkic Computational Linguistics, TurCLing*, pages 1–6.

Lloyd Balderston Swift. 1963. *A reference grammar of Modern Turkish*, volume 19. Indiana University.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The Penn treebank: an overview. In *Treebanks*, pages 5–22. Springer.

Marat M. Yavrumyan, Hrant H. Khachatrian, Anna S. Danielyan, and Gor D. Arakelyan. 2017. ArmTDP: Eastern Armenian Treebank and Dependency Parser. In *XI International Conference on Armenian Linguistics, Abstracts*.

Daniel Zeman, Filip Ginter, Jan Hajič, Joakim Nivre, Martin Popel, Milan Straka, and et al. 2017. 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.

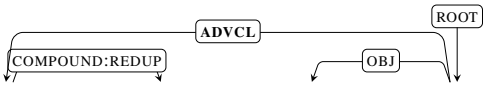
A The Proposed Guidelines for Turkish in the UD Project

For the syntactic analyses and for the annotations, we have accepted most of the already-existing definitions and explanations for the syntactic relations for Turkish in the UD website[2]. Even though, the page itself is in UD version 1.0, the links to the explanations of the syntactic relations are in UD version 2.0. For our analyses, we have edited and/or introduced a total of eight syntactic relations: `advcl`, `advmod`, `compound`, `iobj`, `nmod:poss`, `nsubj`, `obj`, and `obl`. Markdown versions of these guidelines are also available in our github page provided in the paper. In this appendix, we will only include the parts that are different from the original guidelines on the website.

`advcl`

In the explanation of `advcl`, we have included different examples using different morphological inflections to form adverbial clauses. We also included some inflected reduplications as `advcl` as in Example (5).

[2] https://universaldependencies.org/tr/dep/index.html



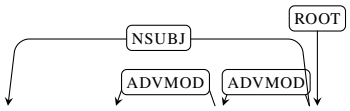
(5) Bilip bilmeden beni suçlama

Bil-ip bil-me-den ben-i suçla-ma.
know-CVB know-NEG-ABL I-ACC blame-NEG

'Don't blame me without knowing anything'

advmod

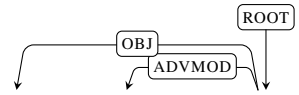
In addition to the explanation and examples, we also included comparative structures with *daha* as in Example (6), adverbs that are formed with a suffix from nouns as in Example (7), and some reduplications as in Example (8).



(6) Ayşe Ali'den daha çevik

Ayşe Ali-den daha çevik.
Ayşe Ali-ABL more agile

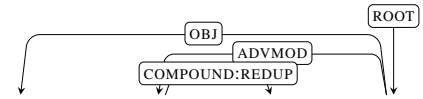
'Ayşe is more agile than Ali.'



(7) Resmi ilgiyle inceliyordu

Resm-i ilgi-yle incel-iyor-du.
picture-ACC attention-COM inspect-PROG-PST

'She was inspecting the picture'



(8) Ödevini zar zor yaptı

Ödev-in-i zar zor yap-tı.
homework-POSS-ACC REDUP difficult do-PST

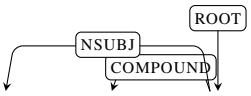
'She struggled doing her homework'

compound

In the guideline of `compounds` we have exampli fied the basic use of the tag as in Example (9). We resort to already-existing guidelines for its use with numbers, and we also used `compound:redup` and `compound:lvc`. However, we have specified the use of the subtype for light verbs, which is `compound:lvc`, and we have limited its use to light verbs that are made up of *et-* and *ol-*. For the rest of the light verbs, we have used `compound` syntactic tag as in Example (10). We also excluded compounds that have syntactic reflex of *-(s)I(n)* from the `compound` tag, instead we have used `nmod:poss` as in Example (11).

(9) çelik yelek

çelik yelek
'steel vest'



(10) annesi hapse girmiş

Anne-si haps-e gir-miş.
mom-POSS prison–DAT enter–EVD

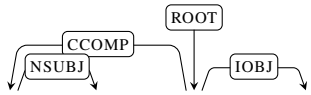
'His mom was put in jail.'

(11) kapı kolu

kapı kolu
'door handle'

iobj

`iobj` is a core nominal argument of the verb apart from the object and subject as in Example (12). Sentences cannot have a `iobj` without having first `obj`.



(12) "Ne bu?" demiş abisine

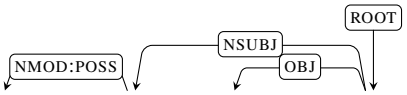
"Ne bu?" de-miş abi-si-ne.
what this say-EVD big.brother-POSS-DAT

'"What is this," he asked to his big brother.'

It is important not to mistake every dative case marked nominal with `iobj` since dative case can be provided semantically and lexically. In those cases, it should be `obl` and `obj`, respectively.

nmod:poss

In our analyses, we also extended the use of NMOD:POSS so that it includes 'X out of Y' constructions for Turkish as in Example (13).



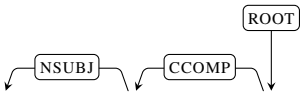
(13) Çocuklardan biri ödevini yapmamış

Çocuk-lar-dan bir-i ödev-in-i
kid-PL-ABL one-POSS homework-POSS-ACC
yap-ma-mış.
do-NEG-EVD

'One of the kids did not do his homework'

nsubj

In addition to the already-existing guidelines, we also specified that the subject of an embedded clause should also be marked with the `nsubj` syntactic tag as in Example (14).



(14) Benim geldiğimi görmüş

Ben-im gel-diğ-im-i gör-müş.
I-GEN come-NMLZ-POSS-ACC see-EVD

'He saw that I have arrived.'

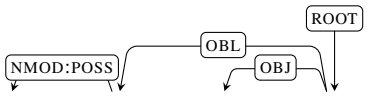
obj

The direct object of a verb is the noun phrase that denotes the entity acted upon.

In Turkish, direct objects typically take either nominative (unmarked), or accusative cases. However, any other case except for genitive can be utilized as well. There are two criteria we use when we decide whether a non-canonically marked object is an `obj` or an `obl`:

- Is the case predictable solely from the semantic denotation of the case?
- Does the verb determine the use of the case?

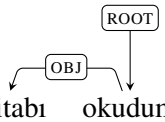
Here, the canonically (marked or unmarked) marked objects:



(15) Hafta sonları kitap okurum

Hafta son-lar-ı kitap oku-r-um.
week end-PL-POSS book read-AOR-1SG

'I read books during weekends.'

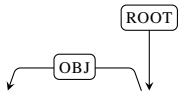


(16) Kitabı okudum

Kitab-ı oku-du-m.
book-ACC read-PST-1SG

'I read the book'.

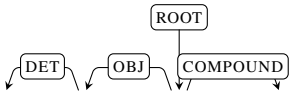
We also utilized the already-existing analyses for partitives and non-case marked noun-phrases. However, we included other non-canonically marked objects as well.



(17) Sana güveniyorum

San-a güven-iyor-um.
you-DAT trust-PROG-1SG

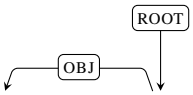
'I trust you.'



(18) Bu evde karar kıldık

Bu ev-de karar kıl-dı-k.
this house-LOC decision do-PST-1PL

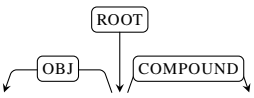
'We have decided on this house.'



(19) Aliden hoşlanıyorum

Ali-den hoşlan-ıyor-um.
Ali-ABL like-PROG-1SG

'I like Ali.'



(20) Çocukla dalga geçiyorum

Çocuk-la dalga geç-iyor-um.
kid-COM wave pass-PROG-1SG

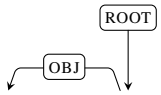
'I am kidding the kid.'

Every case marked noun phrase above is a core element in the sentence and the sentences would be ungrammatical if they were to be left out, thus making them `obj`. This phenomenon is not limited to these verbs only. Many more verbs can utilize non-cannonical object marking in Turkish.

obl

In our syntactic analysis, `obl` relation is used for oblique nominal adjuncts of verbs, adjectives or adverbs. Note that we have used [obj] relation for canonically (accusative and nominative) non-canonically (non-accusative and non-nominative) marked obligatory arguments that are not subjects (objects), and we have used [iobj] relation for core arguments necessitated by the Turkish Grammar.

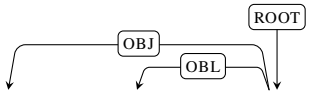
In the examples below, *kitabı* is always the object. However, the other elements that are marked with other cases are adjuncts of the verb and they are not obligatory, which makes them `obl`.



(21) Kitabı okudum

Kitab-ı oku-du-m.
book-ACC read-PST-1SG

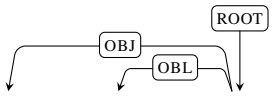
‘I read the book.’



(22) Kitabı çocuklara okudum

Kitab-ı çocuk-lar-a oku-du-m.
book-ACC kid-PL-DAT read-PST-1SG

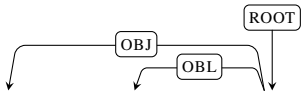
‘I read the book to the children.’



(23) Kitabı uçakta okudum

Kitab-ı uçak-ta oku-du-m.
book-ACC plane-LOC read-PST-1SG

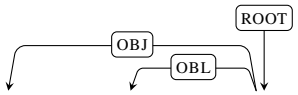
‘I read the book on the plane.’



(24) Kitabı meraktan okudum

Kitab-ı merak-tan oku-du-m.
book-ACC curiosity-ABL read-PST-1SG

‘I read the book out of curiosity.’



(25) Kitabı Ahmetle okudum

Kitab-ı Ahmet-le oku-du-m.
book-ACC PROPN-COM read-PST-1SG

‘I read the book with glasses.’