Improving the Annotations in the Turkish Universal Dependency Treebank

Utku Türk[‡], Furkan Atmaca[‡], Şaziye Betül Özateş[*], 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,
balkiz.ozturk,gungort,arzucan.ozgur@boun.edu.tr

Abstract

This study focuses on a comprehensive analysis and manual re-annotation of the Turkish IMST-UD Treebank, which was automatically converted from the IMST Treebank (Sulubacak et al., 2016b). In accordance with the Universal Dependencies' guidelines and the necessities of Turkish grammar, the existing treebank was revised. The current study presents the revisions that were made alongside the motivations behind the major changes. Moreover, it reports the parsing results of a transition-based dependency parser and a graph-based dependency parser obtained over the previous and updated versions of the treebank. In light of these results, we have observed that the re-annotation of the Turkish IMST-UD treebank improves performance with regards to dependency parsing.

1 Introduction

With its unique set of tags and as a multilingual framework of natural language processing (NLP), the Universal Dependencies (UD) Project[1] offers researchers a common ground regarding the specific features of every language. However, not all languages contribute equally to this project. One language that has not been thoroughly studied is Turkish, which constitutes a canonical example of agglutinative language. Turkish bears a unique challenge in the pursuit of syntactic parsing due to its highly agglutinative morphology and flexible word order patterns. The inefficacy and the inaccuracies of the treebanks previously proposed for Turkish hinder the development of its use in NLP frameworks and its contribution to the UD Project. With this in mind, we set out to find a way to take Turkish Treebanks a step further to overcome the challenges Turkish poses.

In studies that have been previously conducted on Turkish treebanks (Sulubacak et al., 2016a; Sulubacak et al., 2016b; Pamay et al., 2015; Pamay and Eryiğit, 2014; Oflazer et al., 2003; Çöltekin, 2015), there is one very obvious drawback: They differ significantly from one another in terms of their compliance with the rules of Turkish grammar and the UD annotation patterns. While the ITU Validation set and the Turkish PUD treebank follow the finely grained rules of Turkish grammar and annotation guidelines of the UD Project, the IMST-UD Treebank has a coarsely grained structure and inconsistencies resulting from its automated creation. For example, the Turkish PUD treebank utilizes language specific syntactic relations, i.e., `obl:tmod`, `acl:relcl`, `det:predet`, and `flat:name`, quite generously, and the annotation of the Turkish PUD does not have problems that stem from nominalization-based morpho-ortographical similarities (Türk et al., 2019). Our objective is to unify the annotation patterns and decisions based on the requirements of Turkish grammar across the existing treebanks through manual annotations of all the sentences in the Turkish treebanks within the UD framework. By doing so, we aim to create a basis for a future treebank that we will build. The new treebank will consist of an additional ten thousand unique sentences. We have recently re-annotated the Turkish PUD Treebank with our proposed guidelines

[1] https://www.universaldependencies.org

(Türk et al., 2019). In this study, we unify the annotation scheme of the Turkish treebanks and manually re-annotate all the sentences of the IMST-UD Treebank.

2 IMST-UD Treebank

Following the initial efforts to build an annotated treebank, in English (Marcus et al., 1993) and in other languages, Atalay et al. (2003) and Oflazer et al. (2003) introduced a pioneering METU-Sabancı treebank for Turkish. Then, this treebank was re-annotated as IMST Treebank and automatically converted to the UD framework, which resulted in unrivaled scores in NLP tasks for Turkish (Sulubacak et al., 2016a). However, this re-annotation process did not include a linguistics team; instead, it consisted of only one linguist and a team of NLP specialists. Moreover, every annotator worked on their own share of sentences, which resulted in a highly disharmonious picture overall.

As an immediate result of the non-communicative nature of the annotation process, the IMST-UD treebank is incoherent on most of the items, with inconsistencies ranging from ongoing debates on the distinction between `obl` and `obl:arg`[2] in case-marking languages to the very simple concepts of `root` and `punctuation`. Nevertheless, the treebank's morphological segmentation and *inflectional groups* analysis, which refers to the morphosyntactic division of words with respect to their derivational morphology, made the IMST-UD Treebank one of the cornerstones of Turkish and Turkic treebank studies. For this study, we only included the IMST-UD Treebank in the re-annotation process. We have recently re-annotated the Turkish PUD Treebank (Türk et al., 2019) and plan to include the ITU Validation Set in the future with another brand new UD Treebank. We have excluded the Grammar-Book Treebank, created by Çöltekin (2015), as it may contain incomplete phrases and structures that may hinder the parser.

3 The Boğaziçi-ITU-METU-Sabancı Treebank (BIMST)

3.1 Overview

In this work, we have manually examined all the sentences in the IMST-UD treebank with the version UD 2.2 and updated the annotation of the syntactic relations and the head-dependency relations. We have used the most recent version of the treebank that is available on GitHub[3]. For this paper, we have excluded the improvement of the morphological segmentations in the treebank, and accepted the previous morphological parsing (Çöltekin, 2016). The reason for this exclusion is that we believe that the inconsistencies and the annotation problems demanded the most urgent attention. For this reason, we have accepted the same morphological segmentation principles that are used in the most recent version of the treebank.

3.2 Process

In the manual re-annotation process of the IMST-UD Treebank, we first reviewed the definitions of UD (Nivre et al., 2016) and the Stanford Dependencies work (De Marneffe et al., 2014) that influenced important components of the UD framework. Then, we compared our example sentences with the examples in the UD Project website for Turkish and also cross-linguistically. Having settled on the definitions, we reported errors and inconsistencies found in the existing treebanks. The errors and inconsistencies typically resulted from the automated nature of the IMST-UD treebank tags or were due to incorrect linguistic analyses.

To resolve the inconsistencies in the previous treebanks, a team of three linguists and three NLP specialists was formed for the current study. Moreover, discussions were held for potential solutions and their merits according to both Manning's Law (Nivre et al., 2017) and the necessities of Turkish grammar. The criteria we used took into consideration the fine grained distinctions established in Turkish linguistics, typological adequacy, ease of rapid and consistent annotation, ease of understandability, and high accuracy in parsing. These criteria helped us narrow down the list of solutions, and in the end, we decided to implement the most feasible ones.

[2]This change follows from the recent discussion that can be found in `https://universaldependencies.org/workgroups/core.html`.

[3]`https://universaldependencies.org/treebanks/tr_imst/`

All the revisions made in the IMST-UD Treebank were recorded and, then were incorporated into the CONLL-U files using the udpipe package in R (Wijffels et al., 2018). The annotated treebanks, the detailed history of changes made in the annotation process, and our proposed new guidelines are available at `https://github.com/boun-tabi/UD_Turkish-BIMST`.

3.3 Revision

In the IMST-UD Treebank, we re-annotated a total of 5,635 sentences. Table 1 depicts the items that we altered most within the IMST-UD Treebank. In addition to these changes, we introduced 8 previously unused dependency types.

Previous Treebank	Updated Treebank	Number of Alterations
NMOD	ADVCL	666
OBJ	CCOMP	481
NMOD:POSS	NSUBJ	312
OBJ	NSUBJ	276
OBL	IOBJ	194
OBL	OBL:ARG	137

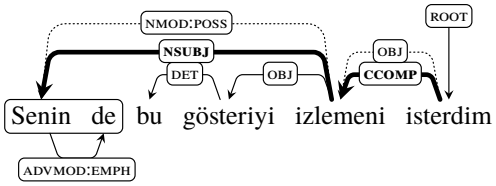
Table 1: The number of alterations that we made for the most frequent changes.

3.3.1 Transparency of Embedded Clauses

The most significant group of changes made in the annotation was based on the inadequate analysis of the internal structure of nominalized embedded clauses in the IMST-UD Treebank. In terms of their external syntax, such clauses behave like regular nouns; in fact, linguistic tests, such as replacement, would deem the whole embedded structure a noun, rather than a clausal object (Göksel and Kerslake, 2005).

However, these structures maintain their predicate's argument structure and inner hierarchy between their own dependents, having the typical properties of a clause in terms of their internal syntax. With this in mind, the outlined genitive phrase was erroneously marked as a possessor in the IMST-UD Treebank as indicated by the dotted lines in sentence (1), even though it does not establish a possessive relation, but syntactically is the genitive marked subject of the nominalized embedded clause (Göksel and Kerslake, 2005). Thus, as shown by the bold line[4] in sentence (1)[5] we marked it as `nsubj`.

(1)



Sen-in de bu gösteri-yi izle-me-ni iste-r-di-m.
you-GEN too this show-ACC watch-NMLZ-POSS want-AOR-PST-1SG

"I would have wanted you to watch this show, as well."

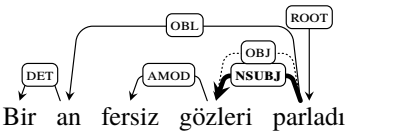
One other issue regarding sentence (1) is that morpho-phonologically identical outputs, the genitive marker on the possessor and the genitive on the subject of embedded sentences, are rendered as possessing the same syntactic function. One other morpho-phonological ambiguity stems from the possessor morpheme in Turkish, which can be ambiguous with the accusative case when it follows a word-final consonant. As seen in sentence (2), *gözleri* is the subject of the unaccusative verb *parla-*. However, it is marked with the suffix *-i* which can be either the possessive suffix on the possessee on nominals or the accusative case. In the IMST-UD Treebank, sentences such as sentence (2) are erroneously marked as `obj` due to the ambiguity mentioned above. However, the ambiguity between the accusative case and the

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

[5]Abbreviations used in this paper are as follows: 1 = first person, ABL = ablative, ACC = accusative, AOR = aorist, CAUS = causative, CVB = converb, DAT = dative, GEN = genitive, NEG = negative, NMLZ = nominalizer, PL = plural, POSS = possessive, PST = past, SG = singular.

possessive suffix is resolved by replacing *gözler* with a noun that has a word-final vowel as in sentence (3). In such environments, the accusative case surfaces as *-yi* as in sentence (1), and the possessive morpheme surfaces as *-si*.

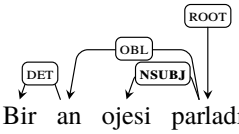
(2)



Bir an fersiz göz-ler-i parla-dı.
one moment dull eye-PL-POSS sparkle-PST

"Her dull eyes sparkled for a moment."

(3)

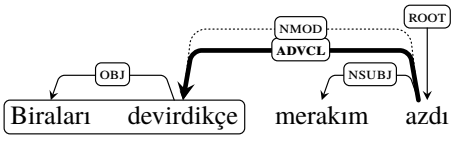


Bir an oje-si parla-dı.
one moment nail.polish-POSS sparkle-PST

"Her nail polish sparkled for a moment."

Another example of the transparency issue within the embedded clauses is the case of converbs and adverbials. In Turkish, these structures also retain their sentential characteristics, which allow them to behave like an embedded clause. This process is not unique to Turkish, and in fact, the English gerund structure is another example of this phenomenon. This is why we, again, ignored the morpho-phonological similarities. Instead of `nmod`, we annotated such items as `advcl`, and the same applied for subjects of embedded structures, which were annotated as `nmod:poss` and we re-annotated them as `nsubj` since they are a core argument for the embedded structure. The new analysis of embedded clauses and core dependents in this section also follows what is proposed by Przepiórkowski and Patejuk (2018) regarding the openness and transparency of an argument structure.

(4)



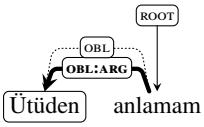
Bira-lar-ı devir-dik-çe merak-ım az-dı.
beer-PL-ACC topple-NMLZ-CVB curiosity-POSS.1SG get.wild-PST

"As I finish my beers, my curiosity peaked."

As indicated by the dotted lines, sentence (4) is also misannotated as `nmod` in the IMST-UD Treebank, even though it is a nominalized converbial structure whose internal argument structure is explicit. We have marked such clauses as `advcl` in accordance with their syntactic function as indicated by the bold line in (4).

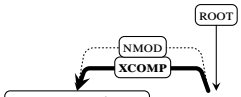
3.3.2 Core vs. Non-core Dependents

Another significant group of changes (*n=139,* n_{total}*=7,899*) made in the re-annotation was based on the definition of core arguments. In addition to canonical object case i.e., accusative case, Turkish also makes use of non-canonical case marking, such as dative, ablative, and locative, for marking obligatory object arguments. The same set of non-canonical case marking can also be used for adjuncts in Turkish. When those cases are selected lexically by the verb, the relation between the nominal head and the verbal head remains the same as if the case on the nominal head was the accusative case. However, non-canonically marked core arguments and adjuncts are both tagged as `obl` in the IMST-UD Treebank. This indicates that non-canonically case-marked arguments, which are obligatory to the sentence, are non-core dependents (De Marneffe et al., 2014). In our analysis, we differentiated the non-core adjuncts from the arguments by following an annotation scheme proposed by Zeman (2017). Zeman (2017) proposes a new syntactic relation for non-canonically marked core arguments, namely `obl:arg`, and differentiates between core objects and oblique arguments, which are also core elements in the sentence but marked with non-accusative case.



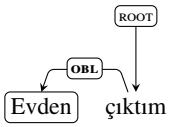
(5) *Ütü-den*
ironing-ABL
anla-ma-m.
understand-NEG-1SG

"I do not know a thing about ironing."



(6) *Düşün-mek-ten* *vazgeç-ti.*
thinking-NMLZ-ABL give.up-PST

"She decided not to think about it."



(7) *Ev-den*
home-ABL
çık-tı-m.
leave-PST-1SG

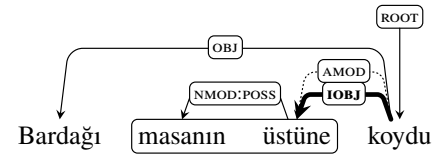
"I have left the home."

In sentences (5) and (6) above, a lexically selected non-canonical case, i.e. ablative, is used to mark core argument dependency. Ablative case, when not lexically selected, is attached to a non-core adjunct and indicates a semantically transparent meaning, such as source, departure, or cause (Erguvanlı-Taylan, 2015) as in sentence (7). However, in sentences like (5) it does not contribute to the meaning of the verb *anlamak* (meaning 'to understand') in any way, showing that it is lexically selected.

In sentence (6) on the other hand, the outlined constituent is not a nominal, but a nominalized non-finite clause, which is again a core argument. Nominalized clauses are formed with a subordinating suffix, as in item (6), but again are marked with the non-canonical ablative case (Göksel and Kerslake, 2005).

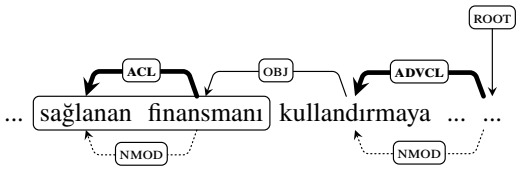
3.3.3 Introduced Dependency Types

In order to increase the linguistic adequacy and the efficiency of NLP tasks, we also introduced dependency types that were previously unused in the IMST-UD Treebank. We propose `advcl` and `iobj` as in sentences (4), (8), and (9), emulating other Turkic treebanks (Tyers et al., 2017). Moreover, we have introduced six more universal syntactic dependencies, namely `xcomp`, `dislocated`, `orphan`, `clf`, `goeswith`, and `dep`.



(8) *Bardağ-ı* *masa-nın* *üst-ün-e* *koy-du.*
glass-ACC table-GEN top-POSS-DAT put-PST

"She put the cup on the table."



(9) *... sağla-nan* *finansman-ı*
... supply-NMLZ financing-ACC
kullan-dır-ma-ya *...*
use-CAUS-NMLZ-DAT ...

"... to make use of the funding provided ..."

In sentence (8), the previous treebank in the UD Project inaccurately marks the relation between the verb and *masanın üstüne* as `amod` since the word group does not modify any noun and, thus, is not an adjectival modifier. One may argue that it is a destination or goal and can be marked as `obl` syntactic relation. However, this analysis would again mean that the word group is not obligatorily required by the event structure of the verb. Hence, we decided to add the indirect object relation to the Turkish Treebank.

4 Experiments

In order to observe the effect of the changes on the parsing performance of the IMST-UD Treebank, a transition-based LSTM dependency parser (Özateş et al., 2018), which is a morphologically enhanced version of Ballesteros et al. (2015) and a state-of-the-art graph-based neural parser (Dozat et al., 2017) are trained on the previous and updated versions of the treebank separately. Both projective and nonprojective dependencies are included in the training and test phases, in contrast to many past studies on the IMST-UD Treebank as well as on its previous versions that used only the projective dependencies (Eryiğit and Oflazer, 2006; Eryiğit et al., 2008; Sulubacak et al., 2016b; Sulubacak et al., 2016a; Sulubacak and Eryiğit, 2018).

[5] 1,022 `advcl` , 351 `iobj` , and a total of 79 for `xcomp`, `dislocated`, `orphan`, `clf`, `goeswith`, and `dep`.

The training part of both versions of the treebank includes 3,685 annotated sentences and the development and test parts include 975 annotated sentences each. That is, we used the original training/development/test partition of the treebank in all of our experiments.

As the pre-trained word vectors, we used the Turkish word embeddings of the CoNLL-17 pre-trained word embeddings from Ginter et al. (2017).

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

4.1 Results and Discussion

Table 2 shows the UAS and LAS F1-scores for words achieved by the parsers on the previous version and the updated version of the IMST-UD Treebank, namely the BIMST Treebank.

		IMST-UD	BIMST
Transition-based parser	**UAS**	65.91	**68.66**
	LAS	59.06	58.98
Graph-based parser	**UAS**	71.55	**75.49**
	LAS	64.86	**65.53**

Table 2: UAS and LAS scores of the two parsers on the previous and updated versions of the IMST-UD treebank.

From the experiment results, we observe that the performances of the parsers in finding correct head-dependent relations increase on the updated version of the IMST-UD Treebank. Although the number of unique dependency tags increased from 33 to 41 with the newly introduced 8 dependency types mentioned in the previous section, the labeled attachment score remains almost the same on the updated version for the transition-based parser and increases for the graph-based parser. Sentence (9) shows an example sub-sentence from the previous version of the treebank and its correct annotation in the updated version. Previously, the dependency tag between *sağlanan* and *finansmanı* was `nmod` although the appropriate tag would be `acl`. The trained parsers predict this dependency tag as `acl`. However, this prediction is counted as false when the previous treebank version is used. In the updated version, such errors were corrected, leading to a better accuracy in terms of parsing of the treebank.

5 Conclusion and Future Work

In this paper, we illustrated the issues in the previous Turkish treebanks, namely the inconsistencies in the annotation and the mismatches between the UD guidelines and the sentences from the treebanks. We also explained our most prominent changes in the re-annotation of the IMST-UD Treebank. We increased the number of syntactic dependency relations that are used to 41, following the UD guidelines more rigidly.

The results demonstrate that the changes made improved the parsing performance with respect to the UAS metric for both of the parsers and the LAS metric for the graph-based parser. Although previously unused dependency relations are included, which pose a challenge for the parser in labeling the dependency relations, there was only a minor decrease in LAS score for the transition-based parser.

As future work, we aim to build the next Turkish treebank on a solid basis. We believe that a new and much more extensive treebank based on a more nuanced and up-to-date corpus should be our trajectory so that Turkish treebanking efforts within the UD framework and NLP processes will yield more accurate results.

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

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*.

Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2015. Improved Transition-based Parsing by Modeling Characters instead of Words with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 349–359. Association for Computational Linguistics.

Çağrı Çöltekin. 2015. A Grammar-Book Treebank of Turkish. In Markus Dickinson, Erhard Hinrichs, Agnieszka Patejuk, and Adam Przepiórkowski, editors, *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*.

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.

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.

Eser Erguvanlı-Taylan. 2015. *The Phonology and Morphology of Turkish*. Boğaziçi Üniversitesi.

Gülşen Eryiğit and Kemal Oflazer. 2006. Statistical Dependency Parsing for Turkish. In *11^{th} Conference of the European Chapter of the Association for Computational Linguistics*.

Gülşen Eryiğit, Joakim Nivre, and Kemal Oflazer. 2008. Dependency Parsing of Turkish. *Computational Linguistics*, 34(3):357–389.

Filip Ginter, Jan Hajič, Juhani Luotolahti, Milan Straka, and Daniel Zeman. 2017. CoNLL 2017 Shared Task - Automatically Annotated Raw Texts and Word Embeddings. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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

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

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666.

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.

Şaziye Betül Özateş, Arzucan Özgür, Tunga Güngör, and Balkız Öztürk. 2018. A Morphology-based Representation Model for LSTM-based Dependency Parsing of Agglutinative Languages. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 238–247.

Tuğba Pamay and Gülşen Eryiğit. 2014. ITU Validation Set for Metu-Sabancı Turkish Treebank. In *Proceedings of the TURKLANG'14 International Conference on Turkic Language Processing*, Istanbul, 06-07 November.

Tuğba Pamay, Umut Sulubacak, Dilara Torunoğlu-Selamet, and Gülşen Eryiğit. 2015. The Annotation Process of the ITU Web Treebank. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 95–101.

Adam Przepiórkowski and Agnieszka Patejuk. 2018. Arguments and Adjuncts in Universal Dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3837–3852.

Umut Sulubacak and Gülşen Eryiğit. 2018. Implementing Universal Dependency, Morphology, and Multiword Expression Annotation Standards for Turkish Language Processing. *Turkish Journal of Electrical Engineering & Computer Sciences*, 26(3):1662–1672.

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

Francis M. Tyers, Jonathan Washington, Çağrı Çöltekin, and Aibek Makazhanov. 2017. An Assessment of Universal Dependency Annotation Guidelines for Turkic Languages. Tatarstan Academy of Sciences, 10.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Abdüllatif Köksal, Balkız Öztürk, Tunga Güngör, and Arzucan Özgür. 2019. Turkish Treebanking: Unifying and Constructing Efforts. In *Proceedings of the 13th Linguistic Annotation Workshop (LAW XIII)*, pages 166–177, Florence, Italy, August 1, 2019.

Jan Wijffels, Milan Straka, and Jana Strakov, 2018. *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit*. BNOSAC, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic.

Daniel Zeman. 2017. Core Arguments in Universal Dependencies. *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 287–296.