DUDU: A Treebank for Ottoman Turkish in UD Style

Enes Yılandiloğlu **Janine Siewert**
University of Helsinki, Finland
{enes.yilandiloglu, janine.siewert}@helsinki.fi

Abstract

This paper introduces a recently released Ottoman Turkish (ota) treebank in Universal Dependencies (UD) style, DUDU. The DUDU Treebank consists of 1,064 automatically annotated and manually corrected sentences. The texts were manually collected from various academic or literary sources available on the Internet. Following preprocessing, the sentences were annotated using a MaCHAMP-based neural network model utilizing the large language model (LLM) architecture and manually corrected. The treebank became publicly available with the 2.14 release, and future steps involve expanding the treebank with more data and refining the annotation scheme. The treebank is the first and only treebank that utilizes the IJMES transliteration alphabet. The treebank not only gives insight on Ottoman Turkish lexically, morphologically, and syntactically, but also provides a small but robust test set for future computational models for Ottoman Turkish.

1 Introduction

Among several treebank projects, the Universal Dependencies treebank project establishing a cross-linguistically consistent treebank annotation scheme for many languages (Nivre et al., 2016, 1659), stands out as the largest collection of treebanks sharing the same annotation scheme (Jøhndal, 2020, 18). Although UD has numerous treebanks for modern languages, historical languages such as Ottoman Turkish remain significantly underrepresented. This paper introduces the DUDU Treebank, one of the first Ottoman Turkish treebanks annotated in the Universal Dependencies (UD) style. The DUDU Treebank consists of 1,064 Latin-transliterated automatically annotated and manually corrected sentences from various genres. The treebank employs the standard Ottoman Turkish transliteration alphabet to handle the alphabet change.

2 Background

Languages from historical periods have always been an engrossing research topic for scholars. The proliferation of computational linguistics methods has accelerated such research in the recent years (e.g., (Harris, 1962)), and UD treebanks project is the manifestation of this process. The UD treebanks aim to provide the sentence's lemma, universal part-of-speech tag (UPOS), XPOS, and mapping for the relationship between arguments (dependency) (see (Nivre et al., 2016) for further explanation). The language analyzed in this paper is Ottoman Turkish, the official and literary language of the Ottoman Empire (Göksel and Kerslake, 2005, 10) and ”a variant of the Perso-Arabic script” consisting of 31 letters (Redhouse, 1884, 1). It was used from the 14th century until the 20th century, up until the decision taken by the Republic of Turkey in 1928 to replace with Latin script (Resmî Gazete, 1928). Unlike the BOUN treebank (Özateş et al., 2024), another treebank for Ottoman Turkish in UD, the DUDU treebank utilizes IJMES Transliteration System to prevent information loss caused by alphabet changes and includes the gender feature, which is absent in modern Turkish but crucial in Ottoman Turkish grammar.

3 Data

A total of 1,064 automatically annotated and manually corrected sentences consisting of 10,012 tokens and 10,287 syntactic words which indicates that 273 tokens are fused forms that are split into multiple syntactic words. The longest sentence has 91 words while the shortest has two

Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025), pages 74–79
March 2, 2025 ©2025 Association for Computational Linguistics

words. The treebank includes 3,133 lemmas and 15 universal POS tags. The morphological annotation covers 67 unique features, including number distinctions (singular: 5,816 instances; plural: 1,001 instances; dual: 3 instances), gender (female: 644 instance; mascular: 110 instances), proper name type (e.g., geography: 173; person: 334), and tense/aspect marking (e.g., past: 1,067 instances; present: 489 instances). Among 38 unique dependency relations, the most common dependency relations are obliques (1,191 instances), noun modifiers (1,291 instances), and objects (657 instances). Various written works from 14th to 20th century were collected as data. Sentences were from various topics including biographical texts, national newspapers, religious texts, fictional works such as stories, instructional texts, popular culture articles, and essays. The main purpose of including data from various registers was to initiate a creation of a representative treebank for the language. The texts were collected from various academic journals, dissertations, and literary sources on the Internet. The texts were transcribed from Perso-Arabic letters to Latin by domain experts; however, with some mistakes. In this research, the Latin transcribed versions were utilized. This initial work focuses on laying the foundation for future research on Ottoman Turkish by leveraging existing modern Turkish treebanks and LLM models instead of focusing on establishing a large treebank.

4 Methodology

In the annotation process, both automatic and manual annotation were leveraged. Initially, we created a seed dataset with only 85 sentences by correcting and manually transforming Ottoman Turkish sentences into their modern equivalents. These sentences were later used to train the annotation model with existing modern Turkish treebanks, as detailed in the following three subsections. Once the initial treebank was created, a model trained on the Ottoman Turkish data was used to annotate unseen sentences without manually transforming phase, which were manually corrected. Following the manual correction phase, these sentences were added into the training dataset and the model retrained. This iterative process significantly improved annotation efficiency.

4.1 Preprocessing

Due to human errors and the lack of standardization in the transcription scheme (e.g., not using a consistent transcription alphabet), a preprocessing step was essential to normalize the data before the annotation phase. This step included comparing the transcribed text with the original Perso-Arabic script manually to correct errors made by the transcriber, if the original script was accessible to the authors. Although the mistakes were minimal, these changes ensured the standardization of the data within the transliteration system for Ottoman Turkish. The primary reason for utilizing the transliteration alphabet instead of modern Turkish alphabet was to more accurately represent Ottoman Turkish with Latin characters. While some transcribers used only the modern Turkish alphabet, some transcribers employed the Ottoman Turkish transliteration alphabet suggested by the IJMES Transliteration System (Cambridge University Press, n.d.), a standardised method for converting the Perso-Arabic script into the Latin alphabet while preventing information loss. In the Ottoman Turkish alphabet, not every letter has a direct equivalent in the modern Turkish Latin alphabet. As a result, multiple Ottoman letters can be represented by the same letter in modern Turkish leading to the loss of information. For instance, the two letters in Perso-Arabic alphabet represented by *k* and *k* in IJMES Transliteration System for Ottoman Turkish are demonstrated by only *k* in modern Turkish alphabet which removes the nuance. This situation, if not addressed with utilizing the IJMES transliteration alphabet, not only leads to morphological ambiguity when words with different meanings are Latinized with the modern Turkish alphabet but also prevents the accurate reflection of Ottoman Turkish orthography. Additionally, it was found that during the transcription phase, punctuation marks were sometimes inserted in the text by the domain expert to make the text clear although there was no punctuation mark in the original sentence in Perso-Arabic script. For such cases, the punctuation marks were removed in preprocessing phase. However, if a word was misspelled in the original text or the punctuation mark was present in the original text, no changes were made. Furthermore, since several books in the data sources were not OCR'd, some sentences were manually transliterated. Following the standardization, the sentences were saved to retrieve

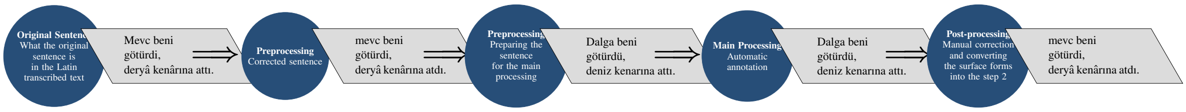
Original Sentenc
What the original sentence is in the Latin transcribed text
Mevc beni götürdi, deryâ kenârına attı.
Preprocessing
Corrected sentence
mevc beni götürdi, deryâ kenârına atdı.
Preprocessing
Preparing the sentence for the main processing
Dalga beni götürdü, deniz kenarına attı.
Main Processing
Automatic annotation
Dalga beni götürdü, deniz kenarına attı.
Post-processing
Manual correction and converting the surface forms into the step 2
mevc beni götürdi, deryâ kenârına atdı.

Figure 1: The initial annotation workflow for Ottoman Turkish treebank creation.

later as ”corrected Latin-transliterated sentences”. Originally, the sentences were quite different from modern Turkish ones, especially lexically. Thus, the words in Ottoman Turkish sentences, which were heavily influenced by Arabic and Persian elements, were manually transformed into their modern Turkish equivalents without altering the morphological structure of the words or the syntactic structure of the sentence since the model was trained via modern Turkish treebanks. This step ensured high accuracy for the model during the automatic annotation process which will be discussed in next subsection. Since the data was in the IJMES transliteration system, LLMs for Arabic and Persian could not be utilized. Moreover, the absence of tools particularly trained on Ottoman Turkish data was another factor to use modern Turkish data to parse Ottoman Turkish sentences.

4.2 Main Processing

After the sentences were manually transformed to resemble modern Turkish, they were ready to be processed by MaCHAMP, ”a flexible toolkit for multi-task learning and fine-tuning of NLP problems”(van der Goot et al., 2021). The MaCHAMP architecture was chosen because of the easiness of the implementation and capability for multi-task learning enabling to annotate all necessary fields for the Ottoman Turkish treebank. The annotation model was trained on over one million sentences from the four existing modern Turkish (tr) treebanks in Universal Dependencies (Sulubacak et al., 2016; Kuzgun et al., 2020, 2021; Marşan et al., 2022). The use of multiple treebanks allowed the model to see more data that enhances its performance in rare and complex linguistic structures. For the task of annotation Ottoman Turkish sentences, we utilised two different transformers. Until we have sufficient data, we used bert-base-multilingual-cased transformer, a large language model to handle multilingual data, (Devlin et al., 2018) as the backbone architecture. After having around 500 sentences, we deployed XLM-RoBERTa base (Conneau et al., 2019), another multilingual transformer model. The model performed five tasks: (I) morphological analysis, (II) lemmatization, (III) UPOS annotation, (IV) XPOS annotation, and (V) dependency parsing. To mitigate overfitting, a dropout rate of 0.2 was utilised and early stopping was applied after 19 epochs. Loss and score results for the training and development phase can be seen in Table 1, below.

Table 1: Model’s performance results.

Task	Train Loss	Development Loss	Train Score	Development Score
Lemmatization	0.3647	1.6630	0.8467	0.6758
Morphological Analysis	0.1216	0.2654	0.9647	0.9350
UPOS	0.2106	0.8509	0.9423	0.8336
XPOS	0.0698	0.4214	0.9731	0.9040
Dependency LAS	0.0460	1.2093	0.9863	0.7965

The model just served to create a base annotation to make the process time efficient and to ease the workload. Afterwards, the automatically annotated sentences were corrected by hand.

4.3 Post-processing

After automatically annotating the transformed sentences via the model, the intermediate transformed sentences were automatically converted back to their forms in the IJMES system using a script. Subsequently, the results were manually reviewed through ”Annotatrix” (Tyers et al., 2018) and corrected. Furthermore, since the model was trained on modern Turkish datasets, it was unable to annotate any feature absent in modern Turkish treebanks such as ”gender”, a significant feature in Ottoman Turkish especially in the construction of noun phrases since all Arabic and Persian words in the noun phrase should share the same gender. Therefore, the ”gender” feature was manually added during the post-processing phase. In addition to the gender feature, the value ”dual” for number feature was also added for words such as *ṭarafeyn* (meaning to ”two sides”), even though it does not exist in modern Turkish. Figure 1 can be used to explain the whole pipeline to create the initial dataset. In Figure 1, following the transcribed sentence, *mevc beni götürdi, deryâ kenârına attı* (the wave carried me, threw to the sea shore), the manually corrected sentence with the IJMES transliteration system can be seen in the second phase. Subsequently, *â* was automatically replaced with *a* and *mevc* was manually converted into *dalga*, its modern counterpart to obtain better performance from the model trained using modern Turkish, not Ottoman Turkish. Following the fourth phase, where the sentence was automatically annotated, the sentence was automatically converted to the second phase, the predicted lemma, and other fields were manually corrected.

4.4 Iterative Training

After establishing the initial treebank with 85 sentences using the method described above, we trained a model using Ottoman Turkish data with MaCHAMP. Subsequently, we annotated Ottoman Turkish sentences with this model and manually corrected the annotations. Then, we retrained the model with more data and used the improved model for the next annotation phase. This process was iterated until we reached 1,064 sentences. During the iterative training phase, we skipped step 3 shown in Figure 1. Furthermore, we found that the XLM-RoBERTa base yielded the best results among various transformer models when sufficient data were available. With the final dataset, the model was trained with a dropout rate of 0.3 and early stopping applied in the 58th epoch. The performance results for the model can be found in the following.

Table 2: Last Model’s Performance Results.

Task	Train Loss	Development Loss	Train Score	Development Score
Lemmatization	0.2509	0.5787	0.9313	0.8534
Morphological Analysis	0.4143	0.9946	0.8898	0.7686
UPOS	0.0389	0.4896	0.9895	0.8976
XPOS	0.0774	0.4927	0.9761	0.8911
Dependency LAS	0.1728	3.5904	0.8933	0.6757

5 Challenges

The main challenge was due to the fact that the Ottoman Turkish language was affected by Arabic and Persian not only lexically, but also grammatically (Göksel and Kerslake, 2005, iii). This meant a comprehensive knowledge of Turkish, as well as Arabic and Persian, was required to address these challenges. The main two challenges related to Arabic elements in Ottoman Turkish were noun phrase structure in Arabic and gender feature. Firstly, dataset contains several Arabic phrases as fixed expressions. Since they function as single units and were mostly idiomatized in Ottoman Turkish, they were treated as single units. A good example of this is *fi’l-vâḳi’*. The phrase, formed with the preposition *fi’* (meaning ”in”) and the noun *vâḳi’* (meaning ”fact”), is in a noun phrase structure and translates to ”in fact” or ”indeed”. Such fixed expressions were shown as single units rather than as separate ones, as shown below:

Table 3: Annotated Arabic fixed noun phrase.

ID	Form	Lemma	POS	Morph	Head	Deprel
-	ve’s-selâm	ve’s-selâm	INTJ	-	30	discourse

On the other hand, we chose to split non-fixed Arabic noun phrases, since each lexical component contributes to the sentence with its morphological and syntactic features, as seen in the *şeyhü’l-beled* (”the religious leader of the town”) example, below:

Table 4: Annotated Arabic non-fixed noun phrase.

ID	Form	Lemma	POS	Morph	Head	Deprel	Misc
4-5	şeyhü’l-beled	-	-	-	-	-	-
4	şeyh	şeyh	NOUN	Case=Nom|Number=Sing|Person=3	5	nmod:poss	-
5	ü’l-beled	beled	NOUN	Case=Gen|Number=Sing|Person=3	6	nmod:poss	-

Another challenge emerged from the gender feature in Arabic and Persian words. Although the gender of words in noun phrases is irrelevant in Turkish, because there is no gender agreement, in Arabic, the words involved in the noun phrase must have the same gender (Göksel and Kerslake, 2005, iii). Gender plays a significant role in Ottoman Turkish noun phrases and the automatic annotation model did not assign the gender feature due to the absence of gender information in the training data. Thus, the gender feature was manually added during the post-processing when necessary. This enrichment aimed to better reflect the linguistic characteristics of Ottoman Turkish in the treebank. Furthermore, Ottoman Turkish, particularly in religious texts, often contains entire sentences in Arabic. To reduce the complexity of the work, such sentences from the treebank were excluded. The challenge related to Persian features in Ottoman Turkish was mainly aroused by *izafet*, ”by which the head of a noun phrase was linked to the modifying noun or adjective that followed it” (Göksel and Kerslake, 2005, iii). In Persian noun phrases, the suffix attaches to the modifier rather than the head noun. If the phrase includes a head noun and an adjective, the suffix applies to the adjective, marking the entire phrase. Although not grammatical in modern Turkish, this structure is common in Ottoman Turkish. An example of *iẓafet* from the dataset can be *ḥuḳûḳ-u meşrû‘* (meaning ”legitimite rights”). In such cases, during the post-processing phase, the morphological analysis of the adjective, which functions as the modifier, was manually corrected as shown below.

Table 5: Example for an Annotated Persian noun phrase.

ID	Form	Lemma	POS	Morph	Head	Deprel	Misc
13	tîr-i	tîr	NOUN	-	16	nmod	-
14	tíze	tíz	ADJ	Case=Dat	13	amod	-

In Table 5, while *tîz* (”sharp”) is an adjective and cannot take the dative case when it modifies a noun, *tîr* (”sword”) in modern Turkish, it is grammatical in Ottoman Turkish. Although challenges listed above both signify the necessity to know, at some degree, the grammar of the languages which were in contact with the target language and demonstrates the requisite of having the post-processing including manual correction to solve the issues.

6 Conclusion and Future Work

To conclude, the DUDU treebank, as the first Ottoman Turkish treebank using the IJMES transliteration alphabet, provides a foundation for further research on different aspects of the Ottoman Turkish language, particularly in lexical, morphological, and syntactic analysis, but also beyond these areas. Furthermore, it also demonstrates that a model trained in the treebanks of a language’s present-day form can be utilized for the analysis of its low-resourced historical form, in this case, Ottoman Turkish leveraging LLM. Future work will focus on expanding the treebank with more data to serve a wide spectrum of language use in Ottoman Turkish and adding new features which modern Turkish lacks; however, Ottoman Turkish has. Unfortunately, for this version, the genres cannot be separated by sentence ids. The order of the sentences is chronology-based rather than genre-based, and the earliest written sentence is at the top. In addition, it is planned to add the original form of the sentence in Perso-Arabic letters to the treebank. Lastly, we plan to publicly release the trained model, which is trained on the final dataset, on the Internet to make it available accessible for further research. In the end, the treebank was created by the first author of this paper with the name DUDU and was published in the UD v2.14 release[1]. The work is currently in progress to expand the treebank to at least 20,000 words for the next release.

References

Cambridge University Press. n.d. *IJMES transliteration chart*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Israa Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (machamp): A toolkit for multi-task learning in nlp. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197. Association for Computational Linguistics.

Aslı Göksel and Celia Kerslake. 2005. *Turkish: A Comprehensive Grammar*, taylor & francis e-library edition. Routledge, London, UK.

Zellig S. Harris. 1962. *String Analysis of Sentence Structure*. Mouton, The Hague.

Marius Jøhndal. 2020. Treebanks for historical languages and scalability. In Elliott Lash, Feng Qiu, and David Stifter, editors, *Morphosyntactic Variation in Medieval Celtic Languages: Corpus-Based Approaches*, pages 15–26. De Gruyter Mouton, Berlin, Boston.

Aslı Kuzgun, Neslihan Cesur, Bilge Nas Arıcan, Merve Özçelik, Büşra Marşan, Neslihan Kara, Deniz Baran Aslan, and Olcay Taner Yıldız. 2020. On building the largest and cross-linguistic turkish dependency corpus. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6. IEEE.

Aslı Kuzgun, Neslihan Cesur, Olcay Taner Yıldız, Oğuzhan Kuyrukçu, Arife Betül Yenice, Bilge Nas Arıcan, and Ezgi Sanıyar. 2021. UD_Turkish-Kenet. `https://github.com/UniversalDependencies/UD_Turkish-Kenet`.

Barış Marşan, Selçuk F. Akkurt, Müge Şen, Melike Gürbüz, Onur Güngör, Şule B. Özateş, Sibel Üsküdarlı, Arzucan Özgür, Tunga Güngör, and Barış

[1] DUDU Treebank is available at (Yılandiloğlu, 2024).

Öztürk. 2022. Enhancements to the boun treebank reflecting the agglutinative nature of turkish. *arXiv preprint*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Şaziye Özateş, Tarık Tıraş, Efe Genç, and Esma Bilgin Tasdemir. 2024. Dependency annotation of Ottoman Turkish with multilingual BERT. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 188–196, St. Julians, Malta. Association for Computational Linguistics.

James W. Redhouse. 1884. *Ottoman Turkish language: A simplified grammar*. The Swiss Bay.

Resmî Gazete. 1928. Türk harflerinin kabul ve tatbiki hakkında kanun (law on the acceptance and application of turkish letters).

Umut Sulubacak, Memduh Gökırmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal dependencies for turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, Osaka, Japan.

Francis M. Tyers, Mariya Sheyanova, and Jonathan North Washington. 2018. Ud annotatrix: An annotation tool for universal dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 10–17.

Enes Yılandiloğlu. 2024. Universal dependencies dudu treebank v2.14. Universal Dependencies, v2.14.