OTA-BOUN: A Historical Turkish Dependency Treebank

Tarık Emre Tıraş[1], **Nureddin Cüneyd Ünal**[1], **Ada Cengiz**[1],
Ece Yurtseven[2], **Esma F. Bilgin Taşdemir**[3], **Şaziye Betül Özateş**[1,*]
[1]Boğaziçi University, Istanbul, Türkiye; [2]Robert College, Istanbul, Türkiye;
[3]Medeniyet University, Istanbul, Türkiye
{tarik.tiras, nureddin.unal, ada.cengiz}@std.bogazici.edu.tr, yurece.27@robcol.k12.tr,
esmabilgin.tasdemir@medeniyet.edu.tr, saziye.ozates@bogazici.edu.tr
*Corresponding author.

Abstract

We present OTA-BOUN v2.0, the largest Universal Dependencies treebank for historical Turkish, consisting of 1,742 manually verified sentences sampled from late Ottoman texts. The annotation process followed a semi-automatic methodology: initial pre-annotation by the UDPipe 2.0 pipeline was refined through manual annotation of dependency relations, part-of-speech tags, and lemmas. A distinctive feature of OTA-BOUN is its dual-script representation: each sentence is provided both in the original Perso-Arabic script and its Latinized transcription, while tokens include aligned forms in both scripts. This dual-layer design enables research on script conversion, cross-lingual transfer, and historical–modern Turkish comparisons. Through detailed analyses on the aforementioned treebank, this study presents a unique and scalable resource, advancing computational studies of historical Turkish and supporting broader efforts in multilingual and diachronic NLP.

Keywords: treebank, universal dependencies, historical Turkish

1. Introduction

The development of robust language resources is fundamental to the advancement of Natural Language Processing (NLP). However, historical languages remain significantly underrepresented, creating a substantial gap in our ability to computationally analyze vast archives of cultural and historical heritage. Although the Universal Dependencies (UD) (Nivre et al., 2016) project has become a standard for creating cross-linguistically consistent treebanks, historical languages are still largely excluded. Ottoman Turkish, the precursor to modern Turkish used for over 600 years, exemplifies this challenge. Following the Turkish language reform movement during the first half of the 20th century, which replaced the Perso-Arabic script with a Latin-based alphabet and reduced the use of Arabic and Persian loanwords, a clear divergence emerged between the two language varieties. This linguistic gap causes modern Turkish NLP tools to be ineffective for processing historical texts, which are characterized by distinct syntactic structures, a different lexicon, and a unique orthography (Öztürk et al., 2025; Aladağ, 2024). The absence of comprehensive treebanks for Ottoman Turkish has hindered systematic computational research, limiting large-scale quantitative analyses of language change and application of NLP to digitized historical documents.

In this paper we introduce OTA-BOUN v2.0,[1] the most comprehensive UD treebank for Ottoman Turkish to date. Expanding significantly on the initial v1.0 release, this version comprises 1,742 manually annotated sentences taken from diverse late Ottoman texts (1880-1928) and nearly 25,000 tokens, making it the largest treebank of its kind. The treebank implements a dual-script representation, providing each sentence in both the original Perso-Arabic script and its Latinized transcription with token-level alignment across scripts. Additionally, manual annotation covers dependency relations, part-of-speech tags, and lemmas following Universal Dependencies guidelines adapted for historical Turkish characteristics. By providing a richly annotated, large-scale and accessible resource, OTA-BOUN v2.0 aims to establish a foundational dataset for the computational study of historical Turkish and to encourage further research in this area.

The remainder of this paper is structured as follows: Section 2 introduces the linguistic background of Ottoman Turkish and reviews related work in the literature. Section 3 details the development of the OTA-BOUN treebank, including data sources, the annotation scheme, and the challenges encountered during dual-script alignment. Section 4 presents a comprehensive error analysis of the pseudo-annotations. Finally, Section 5 concludes the paper and outlines directions for future work.

2. Background and Related work

This section introduces the linguistic background of Ottoman Turkish and highlights its main differences

[1]OTA-BOUN v2.0 is available at https://huggingface.co/datasets/BUCOLIN/ud-ota_boun.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6929–6938
11–16 May 2026.

from modern Turkish. It also presents a brief review of related work in the literature.

2.1. Ottoman Turkish

Ottoman Turkish refers to the historical stage of the Turkish language that was in use from the late 13th century until the early 20th century in the Ottoman Empire. Its writing system was based on an adapted form of the Arabic script, in which five additional characters were introduced to capture Turkish phonemes not represented in the original alphabet. The lexicon of Ottoman Turkish was heavily shaped by contact with Arabic and Persian, incorporating not only extensive loanwords but also complex grammatical constructions from these languages. Such influence is evident, for instance, in the frequent use of long noun phrases, multi-word expressions, and distinctive word-order patterns. As a result, sentences—particularly in earlier Ottoman texts—often exhibit considerable length and structural complexity.

In 1928, the Arabic-based script was replaced by the Latin alphabet as part of the language reform movement initiated in the early Republican period. These reforms also aimed to reshape the vocabulary: a large proportion of Arabic- and Persian-origin words were either replaced by borrowings from European languages or substituted with newly coined native forms. Over time, these reforms produced a marked divergence between modern Turkish and its Ottoman predecessor.

The lexical gap between the two varieties is visible in the distribution of loanwords. Redhouse's well-known Ottoman dictionary A Turkish and English Lexicon (Redhouse, 1884) lists 79,491 entries, 57% of which are of Arabic origin and 12% of Persian origin. By contrast, Moore (Moore and Uni, 2015) reports that among the 3,270 most frequently used words in modern Turkish, only about 25% are Arabic and 6% are Persian. Alternatively, the official Standard Turkish Dictionary published by the Turkish Language Institution (TDK), includes 104,481 entries, of which 6,463 (~6%) are of Arabic origin and 1,374 (~1%) are of Persian origin (TDK, 2005).

Figure 1 presents an example sentence excerpted from *İntikâd*, a collection of letters composed in years 1887-88. The figure displays the image of the portion of the original document containing the sentence (part a) alongside its transcribed version (part b). It further provides the transliteration of the original sentence from the Perso-Arabic script into the Latin alphabet (part c) and its translation into modern Turkish (part d). In parts (c) and (d), Ottoman Turkish words and phrases, together with their modern Turkish equivalents, are highlighted in gray and annotated with corresponding superscripts, demonstrating the lexical and grammatical evolution from Ottoman Turkish to modern Turkish.

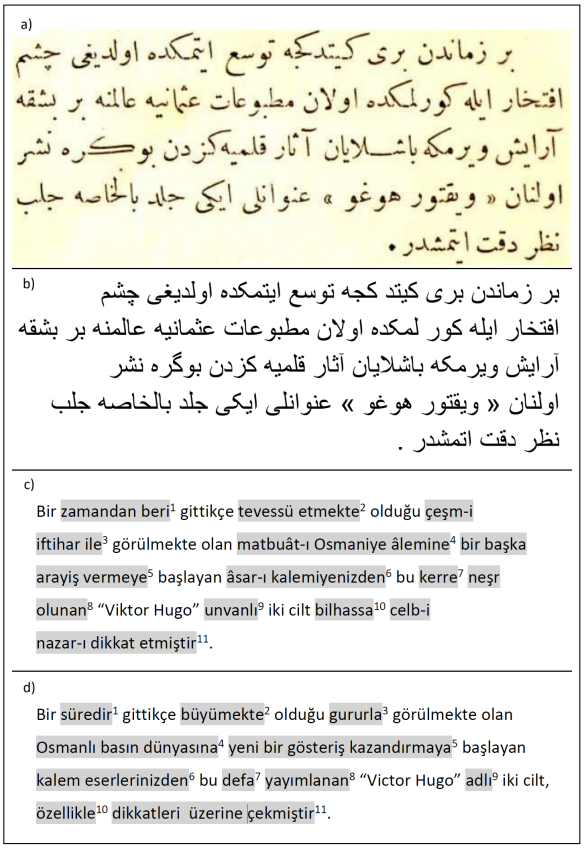


Figure 1: An Ottoman Turkish sentence taken from *İntikâd*, a collection of two writers' correspondence. Part (a) shows the image of the section of the original document containing the sentence. Part (b) presents the transcription of the sentence in digitized form. Part (c) provides the transliteration of the sentence from the original Perso-Arabic script into the Latin script. Part (d) gives the translation of the sentence into modern Turkish. English translation of the sentence: *"Among your literary works that have recently begun to bring a new vitality to the Ottoman press which has for some time been expanding and viewed with growing pride, the two-volume work titled "Victor Hugo" published this time has especially attracted attention."*

2.2. Related Work

The Universal Dependencies (UD) Project[2] is an international collaborative effort to develop cross-linguistically consistent treebank annotations across a wide variety of languages (Nivre et al., 2016). Today, UD includes hundreds of treebanks for over 130 languages, forming the largest publicly available collection of harmonized syntactic

[2] https://universaldependencies.org/introduction.html

resources. These corpora have enabled comparative linguistic studies, cross-lingual parsing experiments, and practical NLP applications. Despite its breadth, however, UD remains skewed towards contemporary, high-resource languages. Historical languages are still severely underrepresented, even though they offer unique insights into diachronic syntax and language change.

Among the few historical languages represented in UD, some of the most established are Ancient Greek and Latin. Large treebanks for these languages were developed more than a decade ago (Bamman and Crane, 2011; Haug and Jøhndal, 2008), covering hundreds of thousands of words from classical texts. They have been used extensively in philological and diachronic research (Futrell et al., 2015; Culbertson et al., 2020; Schlechtweg et al., 2021). Other contributions include treebanks for Gothic, Old Church Slavonic and Old Russian, Old French, and Old English (Haug and Jøhndal, 2008; Berdičevskis and Eckhoff, 2020; Prévost et al., 2024; Martín Arista, 2024), typically derived from historical corpora aligned with Bible translations or early vernacular literature. More recently, resources have also been developed for Classical Chinese (Lee and Kong, 2012), Coptic (Zeldes and Abrams, 2018), and Biblical Hebrew (Swanson and Tyers, 2022). These initiatives demonstrate the value of syntactic annotation for historical texts, yet compared to modern languages they remain relatively rare and often limited in size. Even where multiple treebanks exist, such as for Ancient Greek or Latin, they are tied to specific genres and traditions, limiting broader generalization.

The situation is similar for Turkish. While modern Turkish has several well-established UD treebanks (e.g., Sulubacak et al. (2016); Türk et al. (2022); Kuzgun et al. (2021)) that provide extensive coverage of modern Turkish, none capture the distinctive features of Ottoman Turkish. Ottoman Turkish texts exhibit long, syntactically complex sentences, heavy use of Arabic and Persian lexical and derivational patterns, and discourse connectors rare or absent in modern usage. The absence of a UD treebank for Ottoman Turkish has hindered systematic NLP approaches for the digitized historical Turkish documents and limited comparative studies of language change from Ottoman to modern Turkish. OTA-BOUN v1.0 treebank (Özateş et al., 2025) was developed as a step toward addressing this need, marking the first UD treebank explicitly created for historical Turkish and providing a foundational resource for computational analyses in this domain.

After this first version of OTA-BOUN, the DUDU Treebank (Yılandiloğlu and Siewert, 2025) was introduced as another attempt to model Ottoman Turkish within the UD framework. It includes 1,064 sentences collected from heterogeneous sources such as biographies, newspapers, religious texts, fictional works, and essays. DUDU used the MaCHAMP toolkit for automatic annotation and applied manual correction in iterative cycles. A key methodological choice was to preprocess the data by transforming Ottoman texts into Latinized equivalents using the IJMES transliteration system, without including the original Perso-Arabic script. Some materials had previously been transcribed with modern Turkish orthography, which does not directly correspond with Ottoman Turkish letters and therefore caused a loss of meaning and a need for re-checks on the corpus for consistency. Furthermore, sentences heavily dominated by Arabic words were excluded from the corpus and 'gender' was manually added as a morphological feature during postprocessing, since modern Turkish has a genderless morphology, but Ottoman Turkish reflects Arabic influence.

In comparison with the DUDU Treebank, OTA-BOUN introduces significant improvements in both scale and design. While DUDU contains about 10K tokens, OTA-BOUN includes nearly 25K – making it the largest UD treebank for Ottoman Turkish. A distinctive contribution of OTA-BOUN is its dual-script representation, where each sentence is aligned across Perso-Arabic and Latin transcriptions, supporting both computational and philological analysis. By preserving the original morphosyntactic properties of Arabic- and Persian-derived elements, OTA-BOUN provides a historically informed and computationally usable representation.

3. OTA-BOUN: A UD Treebank for Ottoman Turkish

OTA-BOUN is one of the initial endeavors towards enriching Ottoman Turkish NLP resources and enhancing the effectiveness of existing NLP tools on Ottoman Turkish texts. In this section, we provide information about the data source and language of the treebank, detail our annotation process, and provide treebank statistics. We also outline the annotation challenges with examples and the approaches used to address them.

3.1. Data

OTA-BOUN v2.0 includes 1,742 syntactically annotated sentences written in both original Perso-Arabic script and the Latin-script.

The sentences are sampled from twelve texts by ten different writers. All of the texts are from literature published between 1880 and 1928. There are two articles, excerpts from two history texts, four stories, one memoir, one correspondence, and one excerpt from a novel. Figure 2 visualizes the

sent_id = nut_mk_48
text_arabic = آوروپایه بر هیئت اعزامنه تشبث ایدییور .
text_latin = Avrupa'ya bir heyet i'zâmına teşebbüs ediyor .

1	Avrupa'ya	Avrupa	PROPN	Place	Case=Dat|Number=Sing|Person=3	4	iobj	_	آوروپایه
2	bir	bir	DET	Indef	_	3	det	_	بر
3	heyet	heyet	NOUN	_	Case=Nom|Number=Sing|Person=3	4	obj	_	هیئت
4	i'zâmına	i'zâm	VERB	_	Case=Dat|Number[psor]=Sing|Person=3|Person[psor]=3	5	ccomp	_	اعزامنه
5	teşebbüs	teşebbüs	NOUN	_	Case=Nom|Number=Sing|Person=3	0	root	_	تشبث
6	ediyor	et	VERB	_	Aspect=Imp|Number=Sing|Person=3|Polarity=Pos|Tense=Pres	5	compound:lvc	_	ایدییور
7	.	.	PUNCT	Stop	_	5	punct	_	.

Table 1: Annotation of an example sentence from the OTA-BOUN treebank.

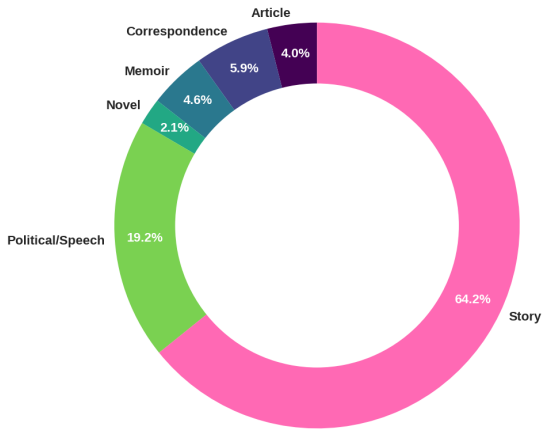


Figure 2: The proportional distribution of genres within the corpus, measured by token count.

proportional distribution of these genres within the treebank, measured by token count. The collection is primarily composed of the *Story* genre, which makes up 64.2% of all tokens, followed by *Political/Speech* at 19.2%. The remainder of the corpus consists of smaller portions from *Correspondence* (5.9%), *Memoir* (4.6%), *Article* (4.0%), and a *Novel* (2.1%). This diverse composition ensures that the treebank captures a range of syntactic and lexical patterns from the period.

In the dataset, we observe various types of lexical replacement between Ottoman and modern Turkish. For example, a single historical term can correspond to a multi-word phrase in contemporary usage, as in *mütarekenâme* versus *ateşkes anlaşması* (armistice agreement).

In addition to lexical differences, we observe syntactic patterns that diverge from modern usage. For example, *dâî-i endişe* literally means "causing anxiety." Here the participle (*dâî*, "causing") precedes its object (*endişe*, "anxiety"), which is the reverse of the typical order in modern Turkish, where the object comes first (*endişe verici*, "anxiety-causing").

3.2. The Annotation Scheme

The annotation process was carried out by two trained linguist annotators, each with substantial expertise in Turkish grammar, general linguistics, and grammatical theory. Their work was supported by two senior computer scientists with extensive background in natural language processing and Ottoman Turkish, who provided guidance throughout the manual annotation phase.

To assess annotation consistency, a randomly selected subset of 50 sentences was double-annotated. For dependency labels, inter-annotator agreement measured on 517 dependency relations reached a Cohen's kappa score of 0.86, with corresponding unlabeled and labeled attachment scores of 88.19% and 81.69%, respectively. For lemmatization, the annotators agreed on 96% of the lemmas. Part-of-speech tagging, being less ambiguous in nature, was performed by a single linguist annotator and subsequently reviewed by the team to ensure internal consistency. The remaining sentences were annotated independently, and upon completion, all annotations were collectively reviewed to resolve any discrepancies.

For the annotation process, we primarily relied on the Universal Dependencies framework, drawing on the Turkish-specific recommendations[3] in cases where additional clarification was required. The present release includes manual annotation of dependency relations, part-of-speech categories, and lemmas, thus providing a richer linguistic layer than the initial version. All material is encoded in CoNLL-U, the standard UD file format, and an example sentence is shown in Table 1. To preserve

[3] https://github.com/boun-tabi/UD_docs/blob/main/_tr/dep/Turkish_deprel_guidelines.pdf

the original script, each sentence contains both Ottoman Turkish form in Arabic script and its Latin transliteration, with the Latinized tokens listed in the second column and the Arabic script forms provided in the final column.

It is important to note, however, that the morphological features included in this release were generated automatically by the pre-annotation pipeline and, unlike the other layers, were not subject to manual verification.

3.3. Treebank Statistics

The initial version of the Ottoman Turkish dependency treebank (Özateş et al., 2025), released with UD v2.14, contains 514 syntactically annotated Ottoman Turkish sentences with a total of 8,794 tokens. With its expansion to approximately 25,000 tokens in the second release, OTA-BOUN ranks as the ninth largest Turkic treebank in the Universal Dependencies collection, according to the comparative overview provided by Akhundjanova et al. (2025).

As summarized in Table 3, the expansion from v1.0 to v2.0 extended the corpus from 514 (8,794) to 1,742 (24,954) sentences (tokens), yielding an approximately threefold increase in size. The average sentence length decreased from 17.1 to 14.3 tokens due to the inclusion of more story-type texts with shorter sentences. The broader range of genres and styles makes the corpus a more representative sample of Ottoman Turkish usage.

Beyond the overall increase in size, OTA-BOUN v2.0 extends the range of annotated dependency relations, as detailed in Table 2. Turkish-specific relations that were absent in v1.0, such as `dep:der`, `discourse:q`, `nmod:part`, `obj:cau`, and `obl:tmod`, are now represented, while relations like `csubj:pass` and `nsubj:pass` occur more frequently. This broader coverage ensures that the treebank better reflects the syntactic variation characteristic of Turkish.

Alongside extending the relation inventory, specific annotation choices were also revised in v2.0. In the initial release, `mark` was limited to complementizers (*e.g., ki, eğer, çünkü*), consistent with earlier practice in several modern Turkish UD treebanks (cf. the comparative table in Türk et al., 2022). In v2.0, this work aligns more closely with the UD guidelines[4]—and with the emerging practice in other modern Turkish treebanks (FrameNet, GB, Kenet, Penn, Tourism)—by extending `mark` to clause-related adpositions and similar subordinators. This update yields a decrease in the relative frequency of `case` (257 → 497; 2.92% → 1.96%) and a corresponding increase in `mark` (27 → 314; 0.31% → 1.24%). Table 4 compares this proportion across several historical and modern Turkish UD treebanks. Notably, our revised `mark` proportion is now comparable to that reported for the DUDU Ottoman Turkish treebank (~1.45%; Yıldiloğlu and Siewert, 2025), indicating a convergent treatment of subordination across Ottoman and modern Turkish UD resources. In addition, many constructions previously annotated as `obl` were reanalysed as `advcl`, following the guideline distinction between oblique dependents and subordinate clauses, which contributes to the higher proportion of `advcl` observed in v2.0.

We also note a substantial increase in the relative frequency of `cop` and `root`. The higher proportion of copular constructions reflects their stronger representation in the expanded dataset, while the increase in root is largely due to the lower average sentence length in v2.0. These shifts affect the overall distribution: the relative share of many other relations appears to decrease, even though their absolute counts continue to rise, and in some cases the proportional growth of certain relations looks more limited than the underlying increase in their counts would suggest.

3.4. Challenges in Annotation

Dependency Annotation. Dependency annotation for Ottoman Turkish is complicated by several linguistic and orthographic factors inherent to the language and its historical context. The main challenges that directly informed the design of our annotation scheme are summarized below:

(i) Complex clause structures: The language frequently exhibits deep syntactic embedding, long noun phrases, and multiple levels of subordination, which complicates the consistent identification of heads and dependents.

(ii) Participial and nominalized modifiers: An extensive use of participles and converbs, often functioning as adjectival or adverbial clauses, makes the consistent application of relations such as `acl` and `advcl` difficult.

(iii) Dual-script orthography: Variation between Perso-Arabic and Latinized forms can obscure syntactic boundaries and introduce inconsistencies in tokenization and alignment.

(iv) Flexible word order and inherited constructions: Ottoman syntax allows for considerable reordering and includes many Arabic- and Persian-influenced multiword expressions that do not align neatly with standard UD conventions.

Lemma Annotation. Lemmatizing historical Turkish requires reconciling diverse scripts, orthographies and etymologies. Many words appear only in Perso-Arabic script, with multiple possible Latin

[4] https://universaldependencies.org/tr/dep/mark.html

Table 2: Counts and percentages of dependency relations in OTA-BOUN v1.0 vs. v2.0.

Relation	v1.0		v2.0	
	Count	%	Count	%
acl	348	3.95	787	3.10
advcl	197	2.24	678	2.67
advmod	396	4.49	1075	4.23
advmod:emph	87	0.99	234	0.92
amod	620	7.04	1681	6.62
appos	2	0.02	2	0.01
aux	39	0.44	181	0.71
case	257	2.92	497	1.96
cc	228	2.59	774	3.05
cc:preconj	12	0.14	58	0.23
ccomp	120	1.36	354	1.40
compound	76	0.86	187	0.74
compound:lvc	246	2.79	676	2.66
compound:redup	33	0.37	95	0.37
conj	607	6.89	1664	6.55
cop	48	0.54	523	2.06
csubj	42	0.48	117	0.46
csubj:pass	–	–	10	0.04
dep	14	0.16	6	0.02
dep:der	–	–	39	0.15
det	508	5.76	1341	5.28
discourse	82	0.93	101	0.40
discourse:q	–	–	71	0.28
dislocated	5	0.06	10	0.04
fixed	6	0.07	25	0.10

First half (25 relations, A–F)

Relation	v1.0		v2.0	
	Count	%	Count	%
flat	87	0.99	228	0.90
flat:foreign	–	–	3	0.01
goeswith	5	0.06	32	0.13
iobj	26	0.30	104	0.41
list	–	–	9	0.04
mark	27	0.31	314	1.24
nmod	137	1.55	215	0.85
nmod:part	–	–	8	0.03
nmod:poss	746	8.47	1995	7.85
nsubj	507	5.75	1433	5.64
nsubj:outer	–	–	3	0.01
nsubj:pass	22	0.25	98	0.39
nummod	57	0.65	199	0.78
obj	557	6.32	1541	6.07
obj:cau	–	–	6	0.02
obl	873	9.91	2412	9.50
obl:agent	4	0.05	12	0.05
obl:cau	–	–	1	0.00
obl:tmod	–	–	194	0.76
orphan	4	0.05	10	0.04
parataxis	10	0.11	56	0.22
punct	1207	13.70	3430	13.50
root	514	5.83	1742	6.86
vocative	7	0.08	52	0.20
xcomp	49	0.56	118	0.46

Second half (25 relations, F–X)

Table 3: Basic statistics of OTA-BOUN: v1.0 vs v2.0

Feature	v1.0	v2.0
Num. of Sentences	514	1,742
Num. of Words	8,812	25,402
Num. of Tokens	8,794	24,954
Avg. Tokens per Sentence	17.10	14.32
Unique POS Tags	16	16
Unique Morphological Features	52	54
Unique Dependency Relations	40	50

Table 4: Proportion of the `mark` relation in historical and modern Turkish UD treebanks.

Treebank	Mark (%)
OTA-BOUN v2.0	1.24
OTA-BOUN v1.0	0.31
DUDU (Yılandiloğlu and Siewert, 2025)	1.45
Atis (Yıldız et al., 2020)	0.02
BOUN (Türk et al., 2022)	0.16
FrameNet (Cirik and Bozşahin, 2015)	0.61
GB (Çöltekin, 2015)	1.16
IMST (Sulubacak et al., 2016)	0.14
Kenet (Çöltekin et al., 2019)	0.36
PUD (Zeman et al., 2017)	0.05
Penn (Oflazer et al., 2018)	0.49
Tourism (Yıldız et al., 2022)	0.26

transliterations. Others are loanwords from Arabic or Persian whose inflectional structure differs from standard Turkish; derivational suffixes may be ambiguous and some words have no direct modern counterpart. This creates annotation dilemmas, as seen with the adverb *ebediyyen* (ابدیاً, 'eternally'). A strict morphological analysis based on its source language, Arabic, would treat the *-en* suffix as a purely inflectional case ending (tanween) and reduce the word to the adjectival lemma *ebedî* ('eternal'). However, within the synchronic system of Ottoman Turkish, this suffix is no longer part of a productive inflectional paradigm but functions as a fixed tool to derive adverbs from Arabic adjectives. Because we treat this as a lexicalized derivational process within Turkish, we chose to lemmatize the word as *ebediyyen* itself, preserving the complete adverbial form.

Since standard morphological analyzers trained on modern Turkish perform poorly on historical vocabulary, we had to verify lemma forms manually using period dictionaries and cross-language lexicons. Even basic segmentation of particles and agglutinative endings are non-trivial when vowels are omitted in the Perso-Arabic script, leading to potential misassignment of lemma boundaries. These factors make automatic lemmatization unreliable. This unreliability is not just a matter of script interpretation but is also rooted in the deep morphophonological alternations of common loanwords. The word for 'age', for instance, derives from Arabic *sinn* (سِنّ), with a geminate final consonant (a doubled consonant sound). This consonant is simplified to sin in its bare form in Turkish, but the original geminate resurfaces when a vowel-initial suffix is added, as in *sinni* ('his age'). This creates a critical ambiguity: the inflected form *sinni* gives no internal clue as to whether its base lemma is the surface form sin or the etymological *sinn*. Since morphological analysis alone cannot resolve this ambiguity, consulting an etymological dictionary—where the latent consonant is recorded—is essential. This reliance on external verification, even for common words, makes lemmatization inherently slow and complex. A more detailed discussion regarding the challenges of Turkish morphosyntax deformation and historical compounds in the initial OTA-BOUN treebank can be found in Özateş et al. (2025).

Table 5: Counts of UPOS and dependency relation corrections by category.

Category	UPOS △	DEPREL △
Article	143	575
Correspondence	126	257
Fiction (novel)	31	87
Fiction (story)	732	1,749
Memoir	141	279
Speech/Political	403	829
All	**1,576**	**3,776**

3.5. Challenges in Dual-Script Alignment

As noted in 3.2, each token in the treebank is represented in both Perso-Arabic script and its Latin transliteration. While this dual-script representation increases the accessibility and linguistic richness of the resource, it also introduces several challenges. A key challenge encountered when adding the original Perso-Arabic script was aligning the tokenized Arabic-scripted versions with their Latin-script counterparts. Orthographic variation in Ottoman Turkish is substantial, with multiple spellings of the same lexical item across texts and authors. Moreover, the correspondence between Perso-Arabic letters and Latin characters is not always straightforward: a single Arabic character may represent different phonemes, or conversely, multiple Arabic graphemes may map onto a single Latinized form. In addition, punctuation and spacing practices in historical texts often obscure token boundaries, complicating tokenization and alignment. For example, in the Arabic-script, forms such as ياپارمی (yapar mı), حيفاكه (hayfa ki), وشواعرده (ve şevâir de) appear as single tokens, whereas in the Latinized form they are separated by spaces. Such occurrences, prevalent in the writing of the conjunctions “de”, “ki” and the question clitic “mı”, create challenges for tokenization and dependency annotation. These discrepancies are particularly common in function words and clitics, while content words generally align more consistently across scripts. Finally, technical issues such as Unicode normalization, diacritic inconsistencies, and bidirectional rendering created further difficulties in maintaining consistency across scripts. Addressing these challenges required a combination of normalization rules, careful manual checking, and the use of historical dictionaries to resolve ambiguous cases.

4. Error Analysis

4.1. Setup and Scope

We followed a semi-automatic pipeline in the annotations: UDPipe 2.0 (Straka, 2018) produced pseudo-annotations for tokenization[5], part-of-speech tags, lemmas, morphological features, and dependency relations, then we manually revised all layers. After that, we conducted an error analysis on the pseudo-annotations and their manual corrections. Nearly half of the sentences contained tokenization or segmentation errors and required correction before manual annotation. To assess labeling errors accurately, we computed confusion statistics only for sentences with identical token alignments in the pseudo- and gold annotations. Hence, the audit was run on a *subset* of the corpus: 1,742 sentences in total, with 1,247 (71.6%) matching between the pseudo annotations and the manual corrections, of which a total of 1,026 (82.3%) sentences were corrected. Table 5 shows universal part-of-speech (UPOS) and dependency relation correction amounts in each treebank category.

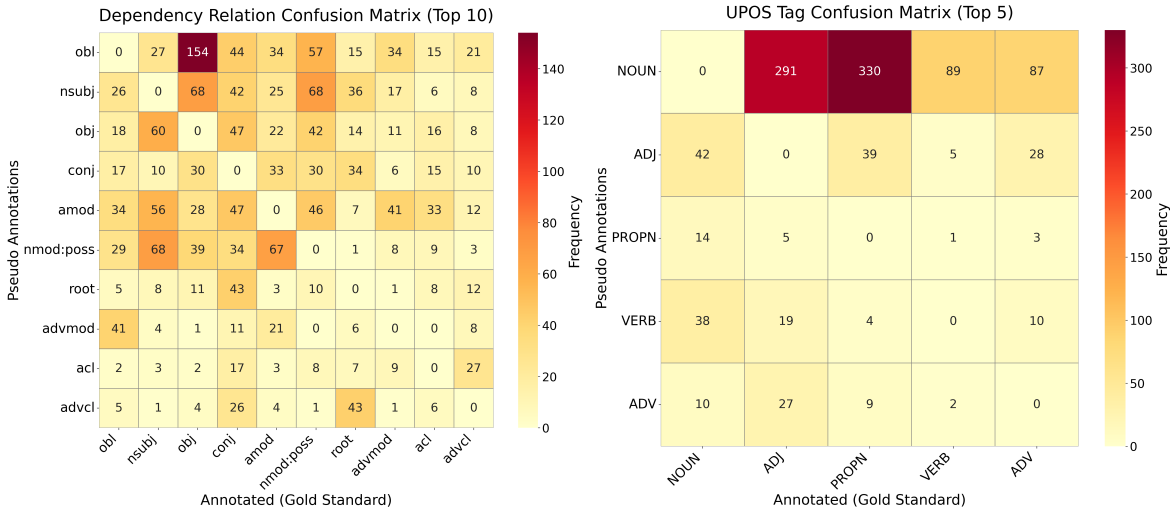
4.2. Confusion Matrices

Figure 3 visualizes the most frequent off-diagonal confusions (rows: UDPipe pseudo; columns: gold) in dependency and POS annotations. Diagonals are suppressed to highlight errors.

Dependency Relations. The high frequency of both `obl` –> `obj` and `nsubj` –> `obj` corrections points to a systematic failure in the model’s valency prediction. Although the parsing model that made the pseudo-annotations was fine-tuned on a small set of Ottoman Turkish data in addition to Modern Turkish, this training data was evidently insufficient for the parser to learn the argument structures of the diverse, low-frequency verbs in the historical lexicon. In Turkish, indefinite direct objects appear in the bare (nominative) case, making them formally indistinguishable from subjects (`nsubj`) or certain obliques (`obl`). Lacking robust lexical statistics for these verbs, the parser defaulted to one of these incorrect labels, necessitating the extensive manual corrections observed in the data.

A second set of systematic errors reveals a core difficulty in processing the complex ’clausality’ of Ottoman Turkish, which relies heavily on nominalized constructions. This is visible in the significant corrections from various modifier and object roles to the subject role, specifically `obj` → `nsubj` (60), `amod` → `nsubj` (56), and `nmod:poss` → `nsubj` (68). This indicates a failure to identify the subjects of embedded clauses; the parser defaults to treating them as simple nominal dependents (`nmod:poss`, `amod`). As noted in the UPOS analysis below, this dependency error often stems from

[5]We used UDPipe only for the tokenization of Latin-scripted sentences. Tokenization of the original Perso-Arabic-scripted sentences was done entirely manually.



(a) Dependency-relation confusions (Top 10).

(b) UPOS tag confusions (Top 5).

Figure 3: Confusion matrices comparing UDPipe pseudo-annotations to OTA-BOUN gold annotations on sentences with matching token counts.

the parser's initial misclassification of the nominalized verb predicate.

A third major pattern, the `nmod:poss` → `amod` correction (67), reveals a systematic failure in parsing borrowed Persian *izāfa* constructions[6]. In Modern Turkish, the parser relies on morphological cues (possessive suffixes) to distinguish `nmod:poss` from `amod`. This cue is absent in the structurally ambiguous Persian *izāfa* (e.g., *hukuk-ı milliye*, which can mean 'national law' (`amod`) or 'the nation's law' (`nmod:poss`)). This ambiguity is critically compounded by a morphosyntactic clash: Turkish case and possessive suffixes attach to the *end* of the head-initial *izāfa* phrase (e.g., *hukuk-ı milliye-**si***), which is the *dependent*, not the *head*. As will be detailed in the UPOS analysis, the parser frequently misclassifies the adjectival dependent (e.g., *milliye*) as a `NOUN`. This combination—a `NOUN` followed by what appears to be another `NOUN` that takes all the Turkish head-marking suffixes—leads the parser to incorrectly default to the head-final Turkish nominal compound structure (`nmod:poss`), necessitating these 67 corrections to `amod`.

UPOS Tags. The most frequent UPOS error is `NOUN`→`PROPN` (330 instances), reflecting a well-known lexical coverage (OOV) issue in historical NLP rather than a grammatical failure. Research on other historical languages, such as Scherrer and Erjavec (2013) for historical Slovene, has shown that a large portion of a historical text's vocabulary—over 20% in their study—is absent from modern dictionaries. This OOV rate stems from the proliferation of historical proper nouns—personal, place, and organizational names—that have fallen out of use. Our 330 `NOUN`→`PROPN` corrections reflect this: the pseudo-annotation model, lacking a lexicon of late-Ottoman proper nouns, defaults to tagging unknown entities as `NOUN`s. While this lexical gap increased annotation workload, it differs fundamentally from systematic structural errors. The following analysis therefore focuses on errors (e.g., `NOUN`→`ADJ`, `NOUN`→`VERB`) revealing the parser's core difficulties with Ottoman Turkish syntax, particularly clausal and adjectival constructions.

A significant UPOS error is the 89 instances of `NOUN` → `VERB` corrections. This is not a simple tagging ambiguity; it is the source of the 'clausality' errors discussed in the dependency analysis. This error occurs when the parser misidentifies a deverbal nominalization or participle—which functions as the predicate (`VERB`) of an embedded clause—as a simple `NOUN`. Once mislabeled, the predicate can no longer, by UD conventions, govern its own subject (`nsubj`), causing a chain of dependency-level corrections (e.g., `nmod:poss` → `nsubj`, `obj` → `nsubj`) noted.

The most significant confusion among nominal categories is the 291 instances of `NOUN` → `ADJ` corrections. This is not a random lexical error; it is a primary driver of the dependency-level failures in parsing Persian *izāfa* constructions, as noted in the analysis of the `nmod:poss` → `amod` error. Because the parser's lexicon lacks many borrowed Arabic and Persian adjectives (e.g., *milliye*), it defaults to tagging them as `NOUN`s. Once this fundamental POS error is made, the parser is faced with an ambiguous `NOUN-ı NOUN` structure. When Turkish suffixes are then added to this final, mis-

[6]This construction connects a noun to a following noun or adjective, typically via a vowel suffix (e.g., *-ı/-i*), to form a complete nominal or adjectival phrase.

classified word (e.g., *hukuk-ı milliye-si*), the parser incorrectly interprets this suffix-bearing "noun" (*milliye*) as the head of the phrase. This single POS failure thus directly triggers the dependency error, forcing the incorrect `nmod:poss` relation.

5. Conclusion and Future Work

This study presented OTA-BOUN v2.0, a comprehensive and manually verified resource for the computational study of historical Turkish. By combining pseudo-annotation with careful manual verification, and providing dual-script representations in both Perso-Arabic and Latinized forms, the treebank facilitates research in script conversion, cross-lingual transfer, and diachronic language analysis. With its scale, quality, and unique design, OTA-BOUN v2.0 not only advances the study of late Ottoman texts but also serves as a valuable asset for multilingual and historical NLP research, opening the way for future developments in these areas.

In the current release, morphological features were generated automatically. While useful as a baseline, they do not fully capture Ottoman Turkish morphosyntactic properties. Future work could add a dedicated annotation layer to represent its rich inflectional and derivational patterns, including Arabic and Persian loanwords. Automatic methods can bootstrap this process, but manual refinement is needed for reliable quality, enabling finer linguistic analysis, cross-linguistic comparison, and more accurate downstream NLP applications.

6. References

Nargiza Akhundjanova, Francis M. Tyers, Nurzhan Sagyndyk, Olimjon Abduraufov, Duygu Ataman, Umut Sulubacak, Daria Kabaeva, Gözde Gül Şahin, Furkan Alperen, and Dilek Küçük. 2025. Parallel Universal Dependencies treebanks for Turkic languages. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, LREC-COLING 2025)*, pages 135–147, Torino, Italy. ELRA and ICCL.

Fatma Aladağ. 2024. Digital humanities and Ottoman studies 2.0. *Journal of Digital Islamicate Research*, 2(1-2):63–89.

David Bamman and Gregory Crane. 2011. The ancient Greek and Latin dependency treebanks. In *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, pages 79–98. Springer.

Aleksandrs Berdičevskis and Hanne Eckhoff. 2020. A diachronic treebank of Russian spanning more than a thousand years. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5251–5256.

Volkan Cirik and Cem Bozşahin. 2015. Framenet-based Universal Dependencies treebank for Turkish. In *Proceedings of the 16th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2015)*. Springer.

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.

Jennifer Culbertson, Marieke Schouwstra, and Simon Kirby. 2020. From the world to word order: Deriving biases in noun phrase order from statistical properties of the world. *Language*, 96(3):696–717.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old Indo-European bible translations. In *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 27–34. Prague.

Aslı Kuzgun, Neslihan Cesur, Olcay Taner Yıldız, Oğuzhan Kuyrukçu, Arife Betül Yenice, Bilge Nas Arıcan, and Ezgi Sanıyar. 2021. UD Turkish Kenet Treebank.

John SY Lee and Yin Hei Kong. 2012. A dependency treebank of classical Chinese poems. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 191–199.

Javier Martín Arista. 2024. Toward a Universal Dependencies treebank of Old English: Representing the morphological relatedness of underivatives. *Languages*, 9(3).

Daniele Moore and Kazuhito Uni. 2015. L'emprunt linguistique comme pont d'apprentissage. quelques réflexions à partir de l'étude des emprunts au Français, à l'Arabe et au Persan dans les langues turques [language loan as a bridge for learning: Reflections from a study on French, Arabic and Persian loanwords in Turkic languages]. *Revue japonaise de didactique du français*, 10:197–213.

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'16)*, pages 1659–1666.

Kemal Oflazer, Cem Keskin, and Çağrı Çöltekin. 2018. The Penn Turkish treebank converted to Universal Dependencies. In *Proceedings of the Universal Dependencies Workshop (UDW 2018)*. Association for Computational Linguistics.

Şaziye Betül Özateş, Tarık Emre Tıraş, Ece Elif Adak, Berat Doğan, Fatih Burak Karagöz, Efe Eren Genç, and Esma F. Bilgin Taşdemir. 2025. Building foundations for natural language processing of historical turkish: Resources and models.

Miraç Göksu Öztürk, Durmus Özkan Sahin, and Erdal Kiliç. 2025. Turkish optical character recognition under the lens: A systematic review of language-specific challenges, dataset scarcity, and open-source limitations. *IEEE Access*.

Sophie Prévost, Loïc Grobol, Mathieu Dehouck, Alexei Lavrentiev, and Serge Heiden. 2024. Profiterole: un corpus morpho-syntaxique et syntaxique de Français médiéval. *Corpus*, (25).

J.W. Redhouse. 1884. *A Turkish and English Lexicon: Shewing in English the Significations of the Turkish Terms*. Number 1. böl. in A Turkish and English Lexicon: Shewing in English the Significations of the Turkish Terms. American mission.

Yves Scherrer and Tomaž Erjavec. 2013. Modernizing historical Slovene words with character-based SMT. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 58–62, Sofia, Bulgaria. Association for Computational Linguistics.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. Dwug: A large resource of diachronic word usage graphs in four languages. *arXiv preprint arXiv:2104.08540*.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Umut Sulubacak, Gülşen Eryiğit, Francis M. Tyers, Eray Yildiz Pamir, and Joakim Nivre. 2016. Universal Dependencies for Turkish. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4140–4146. ELRA.

Daniel Swanson and Francis Tyers. 2022. A universal dependencies treebank of ancient Hebrew. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC'22)*, pages 2353–2361.

TDK. 2005. *Türkçe sözlük*. Number 1. c. in Sözlük Bilim ve Uygulama Kolu yayınları. Atatürk Kültür, Dil ve Tarih Yüksek Kurumu, Türk Dil Kurumu.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Gözde Berk, Seyyit Talha Bedir, Abdullatif Köksal, Balkız Öztürk Başaran, Tunga Güngör, and Arzucan Özgür. 2022. Resources for Turkish dependency parsing: Introducing the BOUN treebank and the BoAT annotation tool. *Language Resources and Evaluation*, 56(1):259–307.

Enes Yılandiloğlu and Janine Siewert. 2025. DUDU: A treebank for Ottoman Turkish in UD style. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 74–79, Tallinn, Estonia. University of Tartu Library, Estonia.

Olcay Taner Yıldız, A. Cüneyd Tantuğ, and Gülşen Eryiğit. 2020. ATIS treebank for Turkish in Universal Dependencies. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*. ELRA.

Olcay Taner Yıldız, A. Cüneyd Tantuğ, and Gülşen Eryiğit. 2022. A Universal Dependencies treebank for Turkish tourism texts. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*. ELRA.

Amir Zeldes and Mitchell Abrams. 2018. The Coptic universal dependency treebank. In *Proceedings of the second workshop on universal dependencies (UDW 2018)*, pages 192–201.

Daniel Zeman, Jan Hajic, Martin Popel, Martin Potthast, and Milan Straka. 2017. Universal Dependencies 2.0: Annotating the PUD treebanks. In *Proceedings of the 15th Conference of the European Chapter of the ACL (EACL 2017)*, pages 55–74. ACL.

Çağrı Çöltekin, Umut Sulubacak, and Francis M. Tyers. 2019. Kenet UD treebank for Turkish: Building a treebank for educational texts. In *Proceedings of the Universal Dependencies Workshop (UDW 2019)*. Association for Computational Linguistics.