DUDU: A Treebank for Ottoman Turkish in UD Style

Enes Yılandiloğlu and Janine Siewert

Department of Digital Humanities, University of Helsinki

Introduction

We present an Ottoman Turkish treebank in UD style, containing 1,064 sentences.

Ottoman Turkish is the official and literary language of the Ottoman Empire [2] and described as ”a variant of the Perso-Arabic script” with 31 letters [3].

Key characteristics:

- **low-resourced** → limited data
- **diachronic evolution** → notable changes over time

Research Questions

How can the modern version of a language be efficiently used to train a model to annotate its low-resource historical variant?

- **Transliteration**: How to preserve information while changing the alphabet?
- **NLP**: How to leverage LLM for annotation task for low resource languages?
- **Language Change**: What are the linguistic differences between Ottoman Turkish and modern Turkish?

Data

- **Data source** → Academic works.
- **Time span** → From 14th to 20th century.
- **Various genres** → News, nonfiction, fiction, government, and poetry
- **Uniform Transliteration** → IJMES Transliteration System [1].
- **Diverse covarage** → 15 POS tags, 53 features, 33 dependencies

Dataset	Tokens	Types	T=L*
Ottoman Turkish	10,012	5,102	4,843

**T=L: % of tokens identical to their lemma.*

Pipeline

1.1 Correct transliteration mistakes ⟹ **1.2** Modernize the lexemes manually + annotate automatically ⟹ **1.3** Back-conversion for lexemes automatically + manual correction ⟹ **2** Train on Ottoman Turkish seed data, annotate and correct more sentences + add them to dataset and retrain the model and expand to 1,064 sentences

Findings

The study demonstrates linguistic aspects that no longer appear in modern Turkish.

Double possessivity

(1) *ḥaleb* ***beglerbegisi***
Aleppo governor-POSS
‘governor of Aleppo’

Unstandardised vowel harmony

(2) *neçün* ***sırruñı*** *fâş* ***itdüñ***
why secret-POSS.ACC revealing make-PST.2SG
‘why did you reveal your secret’

-(n)Hn instead of -(y)Hn

(3) ***ṣunuñ*** *kenârında* *ormanlar* *var*
water-GEN edge-LOC forest-PL exist.PRES.3SG
‘there are forests at the edge of the water’

Results

Below are the performance metrics of the model trained on the full dataset.

Task	Train Loss	Dev Loss	Training Acc.	Dev Acc.
Lemmatization	0.36	1.66	84.67	67.58
Morp. Analysis	0.12	0.26	96.47	93.50
UPOS	0.21	0.21	94.23	83.36
XPOS	0.06	0.06	97.31	90.40
Dependency LAS	0.46	1.20	98.63	79.65

Challenges

- The gender feature in Arabic phrases had to be added manually since the model could not infer it.
- Word order in Persian/Arabic noun phrases differ from its counterpart in modern Turkish which resulted in automatic annotation errors.
- Lack of data transliterated with IJMES.

Conclusions

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.

Our results show that a model trained on the modern form of a language can successfully analyze its low-resource historical variant (in this case, Ottoman Turkish) by leveraging LLM.

Future Works:

- The treebank will be expanded.
- Original forms of the sentences in Perso-Arabic will be added to the treebank.
- The trained model will be publicly available.

QR code for the treebank

[1] Cambridge University Press. *IJMES transliteration chart*, n.d.

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

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