

Unifying the Annotations in Turkic Universal Dependencies Treebanks

Furkan Akkurt¹, Bermet Chontaeva², Çağrı Çöltekin², Mehmet Oguz Derin, Gulgura Dzhumalieva³, Soudabeh Eslami², Tunga Güngör¹, Sardana Ivanova⁴, Murat Jumashev, Aida Kasieva³, Aslı Kuzgun, Büşra Marşan⁵, Balkız Öztürk¹, Chihiro Taguchi⁶, Susan Üsküdarlı¹, Jonathan Washington⁷ and Olcay Taner Yıldız⁸

¹Boğaziçi University

²University of Tübingen

³Kyrgyz-Turkish Manas University

⁴University of Helsinki

⁵Stanford University

⁶University of Notre Dame

⁷Swarthmore College

⁸Özyegin University

Objective

Standardizing annotations in Universal Dependencies (UD) [1] treebanks of Turkic languages.

Introduction

- Treebanks in UD keep growing in number
- Annotation consistency becomes even more important [2]
- 16 treebanks in 8 Turkic languages, 9 in Turkish
- Earlier studies reported consistency issues in Turkic treebanks [3]
- This work started with a workshop (UDTW23, see below) [4]
- Aim: unifying annotations in Turkic treebanks
- Current discussion expected to increase consistency throughout

UDTW23

(UD Turkic Workshop 2023)

- One-day workshop in September 2023 (hybrid) in Istanbul co-located with WG3 meeting
- Beforehand, participants asked to list issues they want to discuss
- 20 exemplary sentences prepared based on issues listed [5]
- Issues discussed through examples since workshop in regular meetings

Information



github.com/ud-turkic

Relevant UniDive working groups

WG1, WG3, and WG4

Issues

- Tokenization

Main source of **tokenization** inconsistency is delineating ‘syntactic words’

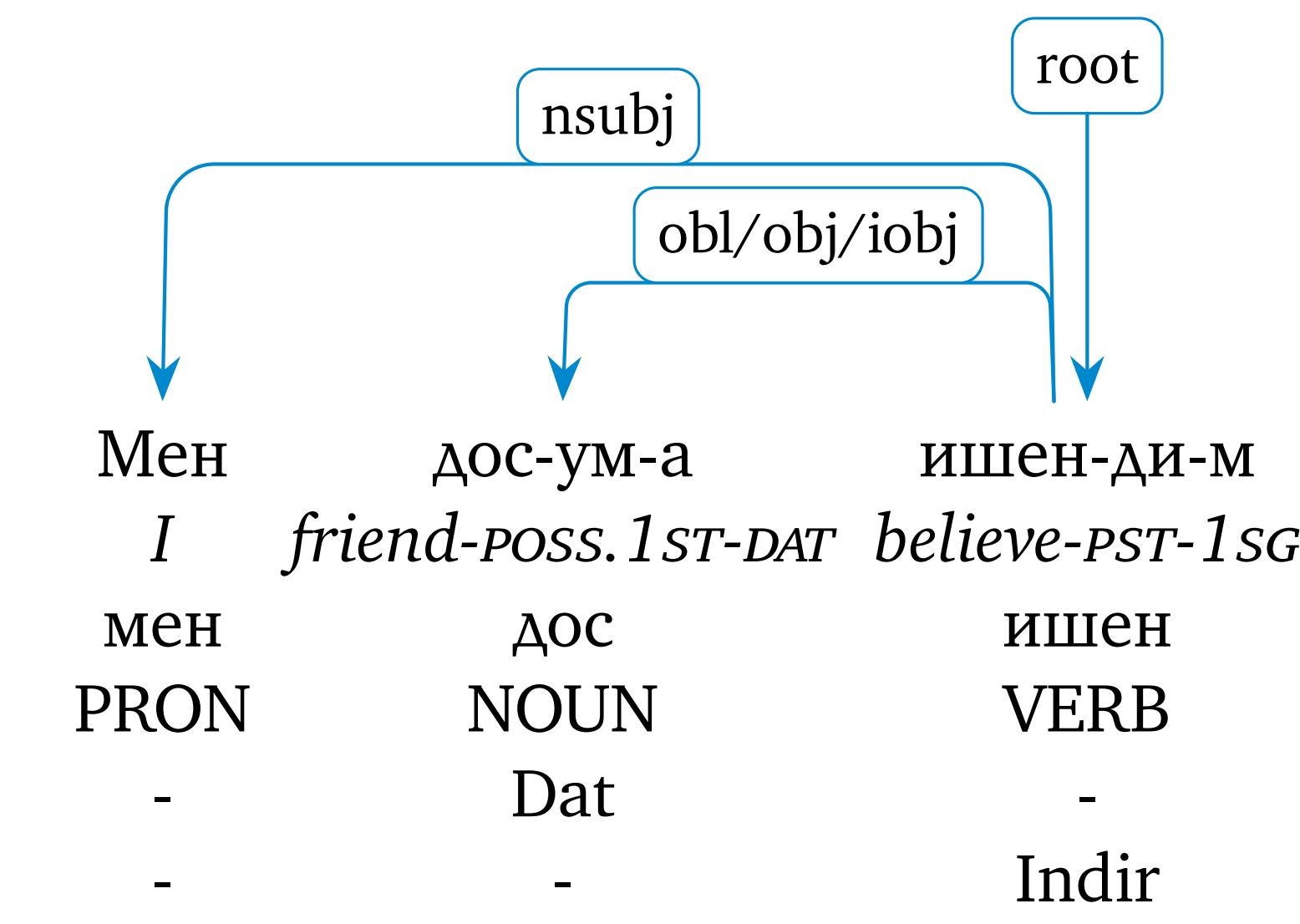
- Morph. feature specification

Many camel-case **tense/mood tags** used inconsistently that should be unified, e.g. GenNecPot ‘general necessitative potential’

- Oblique/object distinction

Non-accusative objects annotated inconsistently:

- obl (**oblique**), based on assessment of event structure
- obj (**object**), based on morphological/syntactic tests (object promotion in passivisation)
- iobj also a possibility



- Question particle

uyu-yor-sun / uyu-yor **mu**-sun
sleep-PROG-P2.SG / sleep-PROG Q-P2.SG
“You’re sleeping”/“Are you sleeping” (tur)

- Can be seen as an *infix*, separated by a space
- Annotations differ on tokenization and assigned POS

- Code switching

Common within Turkic-speaking multilingual communities, with its own challenges in annotation

- Transcription

Coexisting mainstream schemes challenge unified treatment of Translit of MISC attributes

- Pronominalized nouns

- Turkic languages have **pronominal uses of locative and genitive nouns**
- We can imagine 4 different ways of annotation

No segmentation		
amod	orphan/nsubj	root
Büyük	oda-da-ki-ler	uyuyorlar
büyük	room-LOC-ATTR-PL	are sleeping
ADJ	oda	uuu
-	NOUN	VERB
-	Sing/Plur	Plur
-	3	3
-	Loc/Nom	-

“The ones in the big room are sleeping” (tur). Lines below the glosses are for ‘Lemma’, ‘POS’, ‘Number’, ‘Person’, and ‘Case’, respectively.

- one node for two semantic items
- multiple Number and Case features
- Number agreement mismatch
- orphan relation uninformative, nsubj relation misleading

Segmenting before -ki

amod	nmod	
Büyük	oda-da	-ki-ler
büyük	room-LOC	-ATTR-PL
ADJ	oda	PRON
-	NOUN	Plur
-	Sing	3
-	-	Loc
-	-	Nom

- linguistically inaccurate: -ki has attributive function, is not PRON lemma

Segmenting after -ki

amod	nmod	
Büyük	oda-da-ki	-ler
büyük	room-LOC-ATTR	-PL
ADJ	oda	PRON
-	NOUN	Plur
-	Sing	3
-	-	Loc
-	-	Nom

- empty lemma
- empty form when no suffixes

Conclusions

- Workshop served as an important catalyst to foster discussions required to identify and formulate issues on annotation of Turkic languages and developing recommendations for a uniform annotation approach
- Regular discussions still ongoing
- Next step:** discuss all issues in continued meetings with examples from treebanks to reach a collective unification recommendation
- Eventual goal:** write a comprehensive paper detailing issues and decisions to document group’s position
- Study at hand expected to improve overall quality of treebanks and guidelines of UD

References

- [1] universaldependencies.org
- [2] Universalising Latin Universal Dependencies, Gamba et al., 2023.
- [3] An assessment of Universal Dependency annotation guidelines for Turkic languages, Tyers et al., 2017.
- [4] ud-turkic.github.io/udtw23
- [5] github.com/ud-turkic/udtw23/wiki/selected-20-turkish-sentences

Acknowledgements

We thank UniDive, the COST Action CA21167, and Istanbul Technical University for supporting the organization of the workshop in September 2023 that made the present work and future collaboration towards a unified annotation of Turkic UD treebanks possible.

Funded by the European Union



cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

UniDive

FRIEDRICH KARLS
UNIVERSITÄT
TÜBINGEN

ÖZYEĞİN
UNİVERSİTESİ

BOĞAZICI
UNIVERSİTESİ
1963

HELSINKI
YÜLÇÜSTÜ
UNIVERSİTESİ

ELAZIĞ
UNİVERSİTESİ

HELSINKI
YÜLÇÜSTÜ
UNIVERSİTESİ

İSTANBUL
TEKNİK
UNİVERSİTESİ

SWARTHMORE