Unifying the Annotations in Turkic Universal Dependencies Treebanks

Furkan Akkurt[1], Bermet Chontaeva[2], Çağrı Çöltekin[2], Mehmet Oguz Derin, Gulnura Dzhumalieva[3], Soudabeh Eslami[2], Tunga Güngör[1], Sardana Ivanova[4], Murat Jumashev, Aida Kasieva[3], Aslı Kuzgun, Büşra Marşan[5], Balkız Öztürk[1], Chihiro Taguchi[6], Susan Üsküdarlı[1], Jonathan Washington[7] and Olcay Taner Yıldız[8]

[1]Boğaziçi University [2]University of Tübingen [3]Kyrgyz-Turkish Manas University [4]University of Helsinki [5]Stanford University
[6]University of Notre Dame [7]Swarthmore College [8]Özyeğin University

2nd UniDive Workshop, Naples, Feb 2024

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'

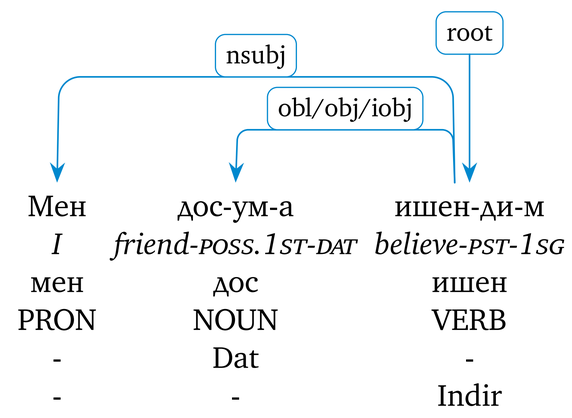
– Morph. feature specification

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



"I believed my friend" (`kir`). Lines below the glosses are for 'Lemma', 'POS', 'Case', and 'Subcat', respectively.

– 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

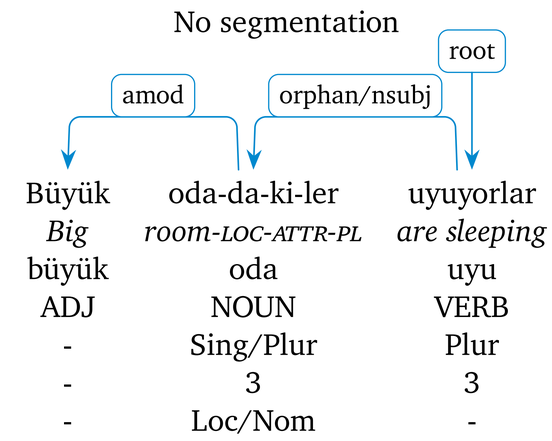
Common within Turkic-speaking multilingual communities, with its own challenges in annotation

– Transcription

Coexisting mainstream schemes challenge unified treatment of `Translit` of `MISC` attributes

– Pronominalized nouns

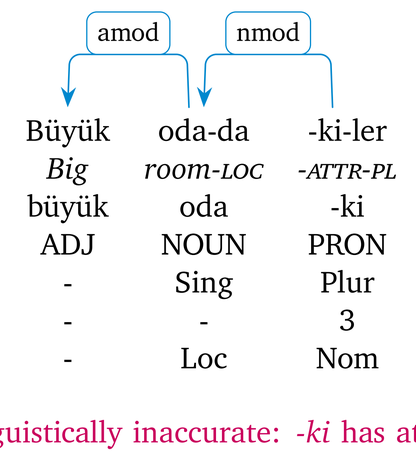
- Turkic languages have **pronominal** uses of **locative and genitive nouns**
- We can imagine 4 different ways of annotation



"The ones in the big room are sleeping" (`tur`). Lines below the glosses are for 'Lemma', 'POS', 'Number', 'Person', and 'Case', respectively.

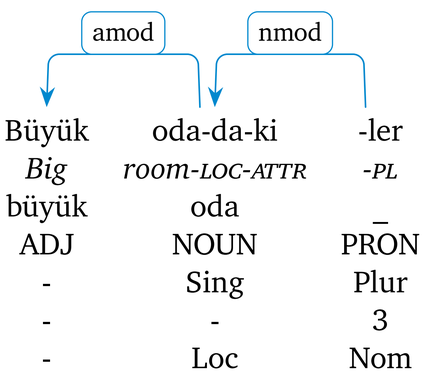
- one node for two semantic items
- multiple `Number` and `Case` features
- `Number` agreement mismatch
- `orphan` relation uninformative, `nsubj` relation misleading

Segmenting before *-ki*



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

Segmenting after *-ki*



- 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.