SYNTAXFEST
Ljubljana
August 26-29 2025

UniDive

Parallel Universal Dependencies Treebanks for Turkic Languages

Arofat Akhundjanova[1], Furkan Akkurt[2], Bermet Chontaeva[3], Soudabeh Eslami[3], Çağrı Çöltekin[3]

[1]Independent Researcher, [2]Boğaziçi University, [3]University of Tübingen

Background

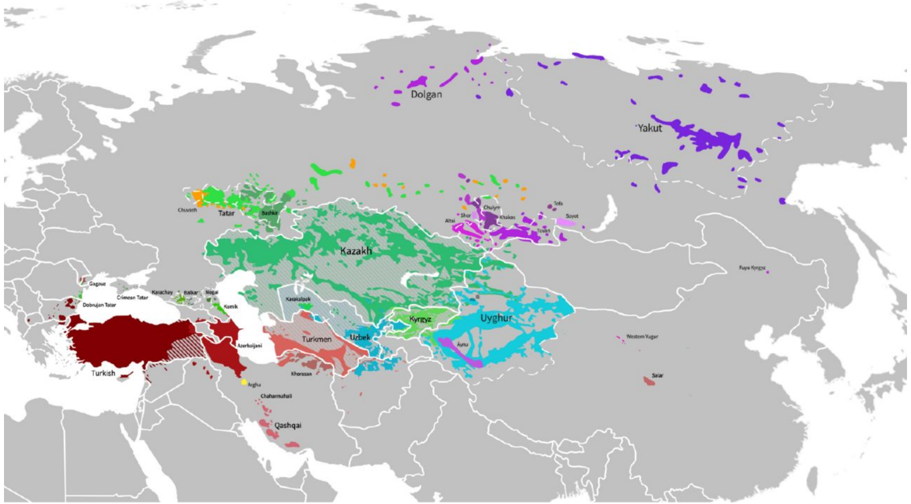
Limited Resources for Turkic Languages

150+ languages in UD framework, but Turkic representation is limited

24 treebanks for 11 Turkic languages with varying quality and size

Annotation inconsistencies across existing treebanks

Very few parallel treebanks for systematic cross-linguistic comparison



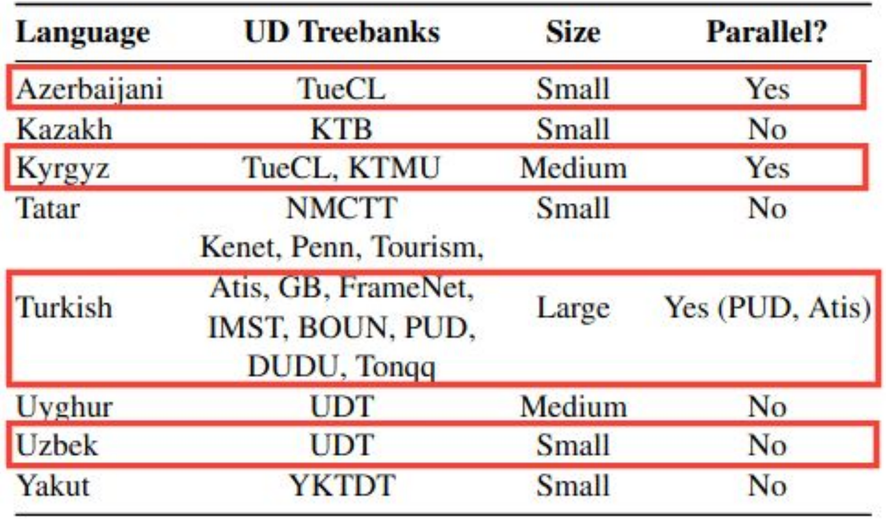
Gap: Despite shared typological features and historical ties, systematic cross-linguistic studies of Turkic syntax are severely limited by lack of parallel annotated data.

Target Languages

Four languages representing three major Turkic branches: Oghuz (Azerbaijani, Turkish), Kipchak (Kyrgyz), Karluk (Uzbek).

Challenges in Turkic UD treebanks:

- **Agglutinative morphology** → long, complex word forms.
- **Complex verb constructions** → serial verbs, auxiliary chains.
- **Low-resource tools** for Azerbaijani & Kyrgyz → limited parsing & corpus development.



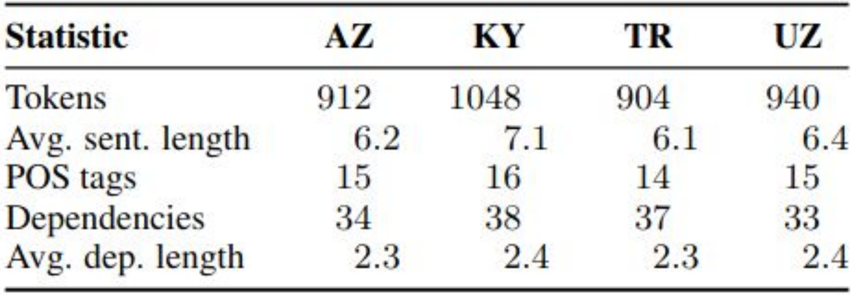
Language	UD Treebanks	Size	Parallel?
Azerbaijani	TueCL	Small	Yes
Kazakh	KTB	Small	No
Kyrgyz	TueCL, KTMU	Medium	Yes
Tatar	NMCTT	Small	No
Turkish	Kenet, Penn, Tourism, Atis, GB, FrameNet, IMST, BOUN, PUD, DUDU, Tonqq	Large	Yes (PUD, Atis)
Uyghur	UDT	Medium	No
Uzbek	UDT	Small	No
Yakut	YKTDT	Small	No

Status of UD treebanks for Turkic languages as of version 2.15.

Dataset Overview

Curated collection of 148 sentences, compiled from multiple sources:

- Cairo corpus: 20 sentences
- UDTW23 corpus: 20 sentences
- Custom examples: 108 sentences illustrating specific grammatical phenomena
- **Strategic Selection:** Sentences chosen to highlight morphosyntactically rich and typologically significant constructions, e.g., pro-drop, auxiliary chains, and non-canonical word orders

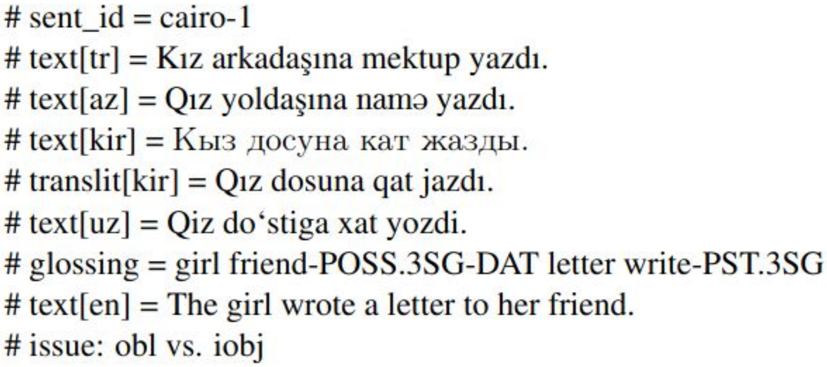


Statistic	AZ	KY	TR	UZ
Tokens	912	1048	904	940
Avg. sent. length	6.2	7.1	6.1	6.4
POS tags	15	16	14	15
Dependencies	34	38	37	33
Avg. dep. length	2.3	2.4	2.3	2.4

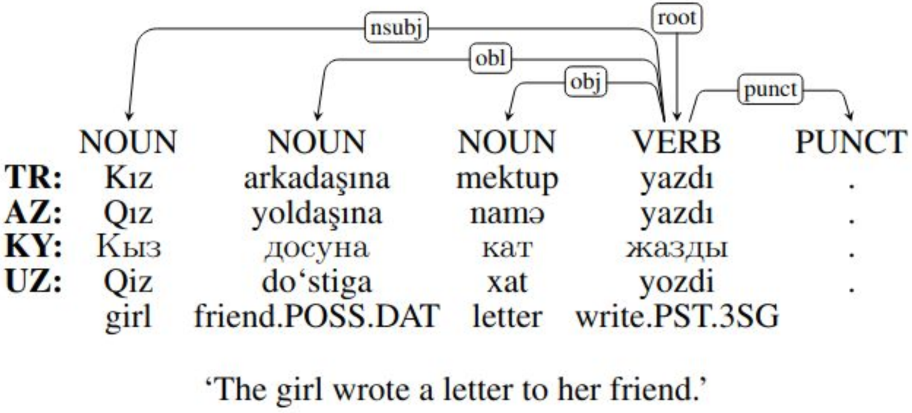
Source Data

Most of the source sentences originate in **Turkish** and were **manually translated** into other languages.

Language script: **Latin** for Azerbaijani, Turkish, and Uzbek; **Cyrillic** for Kyrgyz with transliteration and interlinear glosses provided in the metadata.



```
# sent_id = cairo-1
# text[tr] = Kız arkadaşına mektup yazdı.
# text[az] = Qız yoldaşına namə yazdı.
# text[kir] = Кыз досуна кат жазды.
# translit[kir] = Qız dosuna qat jazdı.
# text[uz] = Qiz do‘stiga xat yozdi.
# glossing = girl friend-POSS.3SG-DAT letter write-PST.3SG
# text[en] = The girl wrote a letter to her friend.
# issue: obl vs. iobj
```



	NOUN	NOUN	NOUN	VERB	PUNCT
TR:	Kız	arkadaşına	mektup	yazdı	.
AZ:	Qız	yoldaşına	namə	yazdı	.
KY:	Кыз	досуна	кат	жазды	.
UZ:	Qiz	do‘stiga	xat	yozdi	.
	girl	friend.POSS.DAT	letter	write.PST.3SG	

‘The girl wrote a letter to her friend.’

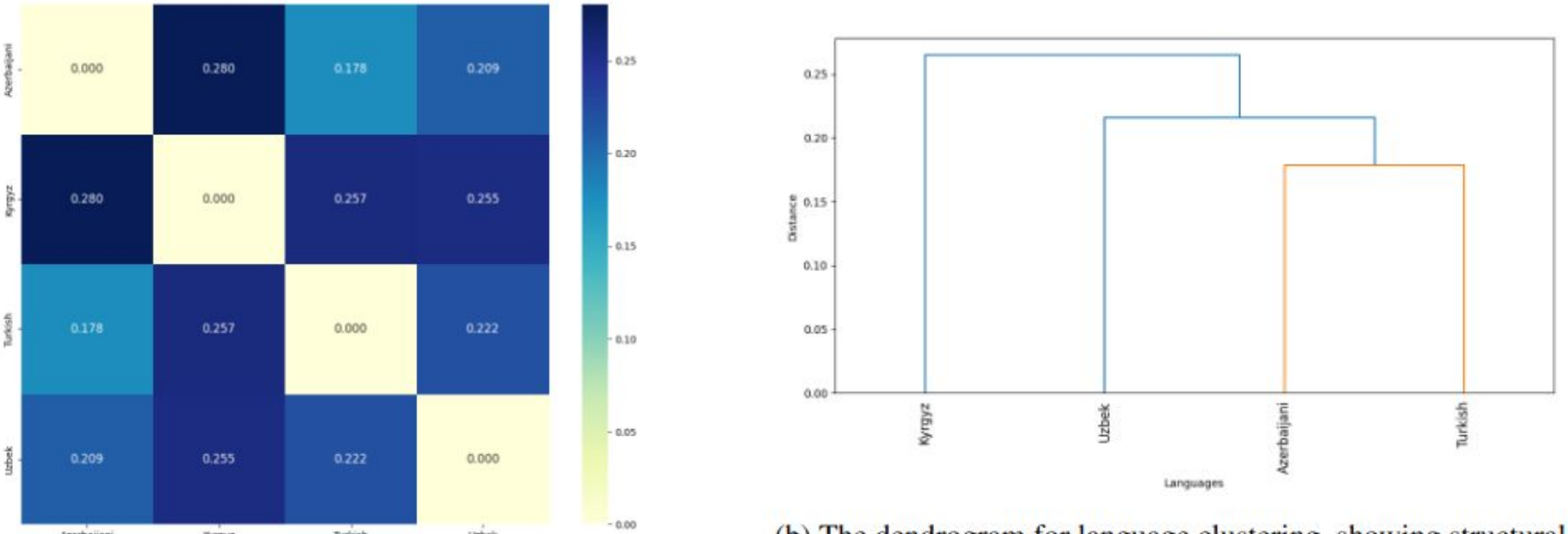
Annotation sample of parallel sentences

Annotation Methodology

- **Hybrid process**: automated processing + manual annotation & revision + expert discussions (linguists & UD experts)
- **Azerbaijani & Kyrgyz TueCL**: extended with new grammar examples & morphological features
- **Turkish**: two parallel strategies → (1) fully manual, (2) automatic (Claude 3.5 Sonnet, 2025) + manual correction → merged results
- **Uzbek**: automated tokenization (NLTK), all other layers annotated manually

Quantitative Analysis

Normalized edit distances based on POS sequences confirm typological relationships



(a) Normalized edit distances based on POS sequences.

(b) The dendrogram for language clustering, showing structural similarities among the languages.

Language-Specific Features

Turkish: Flexible placement of **question particle *mi*** (focus-shifting); determiner-adjective ordering variation

Azerbaijani: Can form **intonation-based questions without particles**.

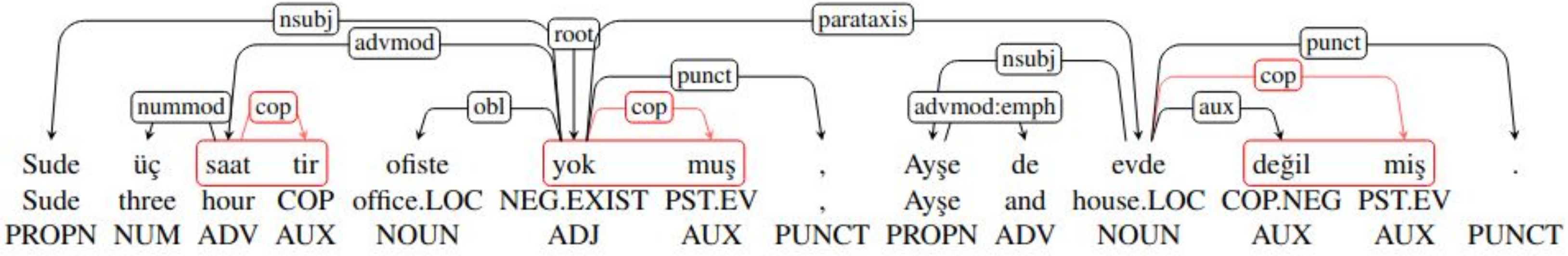
Kyrgyz: Uses **posture/locational verbs** as progressive auxiliaries; can form **compound nouns without possessive suffixes**.

Uzbek: Longest dependency lengths

Annotation Challenges

Copular constructions

- Challenge: copula realized as affix → inconsistent analyses
- Solution: treat copular affixes as **AUX** with **cop** relation to main predicate

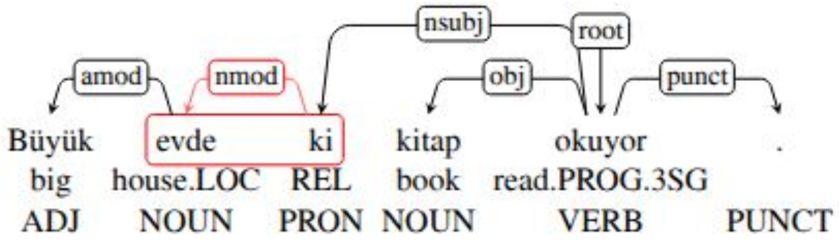

nsubj
advmod
root
parataxis
punct
nsubj
punct
nummod
cop
obl
cop
advmod:emph
aux
cop
Sude üç saat tir ofiste yok muş , Ayşe de evde değil miş .
Sude three hour COP office.LOC NEG.EXIST PST.EV , Ayşe and house.LOC COP.NEG PST.EV
PROPN NUM ADV AUX NOUN ADJ AUX PUNCT PROPN ADV NOUN AUX AUX PUNCT

‘Sude was not at the office for three hours and Ayşe was not at home.’

Annotation Challenges

Pronominalized locatives (-ki) (Washington et al., 2024)

- Challenge: complex genitive/locative forms, hard to auto-annotate
- Solution: treat **-ki** as separate subtokens → preserves full linguistic info



‘The one in the big house is reading (a) book.’

Concluding Remarks

- **First aligned UD treebanks** for 4 Turkic languages → foundation for comparative studies & cross-lingual NLP
- **Limitations**: small size, constructed examples, focus on written/formal registers
- **Future directions**: expand texts & languages, analyze more morphosyntactic phenomena, invite community collaboration
- **Takeaway**: valuable starting point, demonstrating feasibility & paving way for broader Turkic resources
- **Acknowledgments**: Turkic UD working group & COST Action CA21167 (UniDive)
- 🔗 **Collaboration Welcome!**
  Join the Turkic UD working group to expand and improve these resources

Resource Availability

📂 Universal Dependencies v2.16

All treebanks publicly available as part of the official UD release

Treebank Names

- UD_Azerbaijani-TueCL
- UD_Kyrgyz-TueCL
- UD_Turkish-TueCL
- UD_Uzbek-TueCL

THANK YOU!

ANY QUESTIONS?