An assessment of Universal Dependency annotation guidelines for Turkic languages

Francis M. Tyers[a], Jonathan Washington[b],
Çağrı Çöltekin[c], and Aibek Makazhanov[d]

(a) Школа лингвистики, Высшая школа экономики, Москва;
(b) Linguistics Department, Swarthmore College, Swarthmore;
(c) Seminar für Sprachwissenschaft, Universität Tübingen;
(d) National Laboratory Astana, Nazarbayev University, Astana

Overview of talk

- Universal Dependencies: what & why
- Universal Dependency annotated treebanks for Turkic languages (Kazakh, Turkish, Uyghur)
- Differences between the treebanks
- Parsing performance
- Open questions
- Conclusion

Universal dependencies (UD)

What UD is

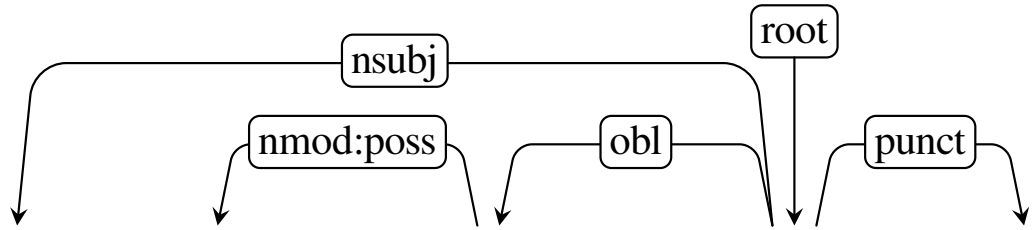
A set of guidelines for syntactic & morphological annotation of text

What UD offers

- Agreed-upon "universal" / unified tag sets (for any language)
  - Part of speech
  - Morphological features
  - Dependency (syntactic) relations
- Support
  - Guidelines for use of the tags
  - An active community which can assist with difficult use cases
- A venue for publication of
  - language-specific annotation guidelines
  - annotated open-source text corpora
- Corpora (=usage examples) in a lot of languages (always growing)

Universal dependencies (UD)

Demonstration



	Мин	Казан	шәһәренә	килдем	.
Gloss	*I*	*Kazan*	*to the city of*	*I came*	
POS	PRON	PROPN	NOUN	VERB	PUNCT
Lemma	мин	Казан	шәһәр	кил	-
Number	Sing	Sing	Sing	Sing	-
Case	Nom	Nom	Dat	-	-
Person	1	3	3	1	-
Number[psor]	-	-	Sing	-	-
Person[psor]	-	-	3	-	-
VerbForm	-	-	-	-	-
Tense	-	-	-	Past	-
Evident	-	-	-	Fh	-

Turkic languages in UD

Current status

Large treebanks in three Turkic languages

- Kazakh
- Turkish
- Uyghur

The full list:

Treebank	Language	Sentences	Words	Annotation	Genre
Kazakh-UD	Kazakh	1047	10 032	manual annotation	Wikipedia, fiction
IMST-UD	Turkish	4660	48 093	semi-auto. conversion	news, social media
Turkish-PUD	Turkish	1000	16 886	auto./manual annotation	translated news
Turkish-GK	Turkish	2803	17 800	manual annotation	grammar examples
Uyghur-UD	Uyghur	100	1662	semi-auto. conversion	fiction

Turkic languages in UD

Turkish treebanks

- IMST-UD treebank (sulubacak2016) ← IMST treebank (sulubucak2016imst) ← METU-Sabancı (oflazer2003)
- main treebank: Turkish-PUD
- Turkish-GK (coltekin2015tlt) UD v1.3, grammar book examples

Turkic languages in UD

Kazakh treebank

- 1 treebank, 1109 trees, 10894 tokens
- Tyers & Washington (2015), Makazhanov (2015) [TurkLang!]
- Tokenisation per Apertium standards
- Mostly compatible with UD v2.0

Turkic languages in UD

Uyghur treebank

- Converted from Uyghur treebank (aili2016)
- Contains surface forms, POS, and dependency relations
- Does not contain lemmas or morphological features

Turkic languages in UD

Other Turkic treebanks

- Tuvan (Ageeva and Tyers, 2016), approx. 1000 tokens;
- Crimean Tatar (Ageeva and Tyers, 2016), approx. 1000 tokens.

Turkic languages in UD

Differences between the treebanks: part-of-speech tagging

Defective pronouns or adverbs?

Annotation in current corpora:

language	word	gloss	POS	dep rel
Turkish	nerede	*where*	`PRON`	`obl`
Turkish	nereden	*from where*	`PRON`	`obl`
Kazakh	қайда	*where*	`ADV`	`advmod`
Kazakh	қайдан	*from where*	`ADV`	`advmod`

Analysis as pronouns

- in Turkish they appear to be pronouns with all case forms

Analysis as adverbs

- in Kazakh they don't appear in most cases (nom, gen, etc.)

Turkic languages in UD

Differences between the treebanks: morphological features

- Turkish: `Person=3` for any nominal
  Kazakh: not marked
- Turkish: `Polarity=Pos/Neg`
  Kazakh: only `Polarity=Neg` marked
- no morphological features in Uyghur corpus

Turkic languages in UD

Differences between the treebanks: tokenisation

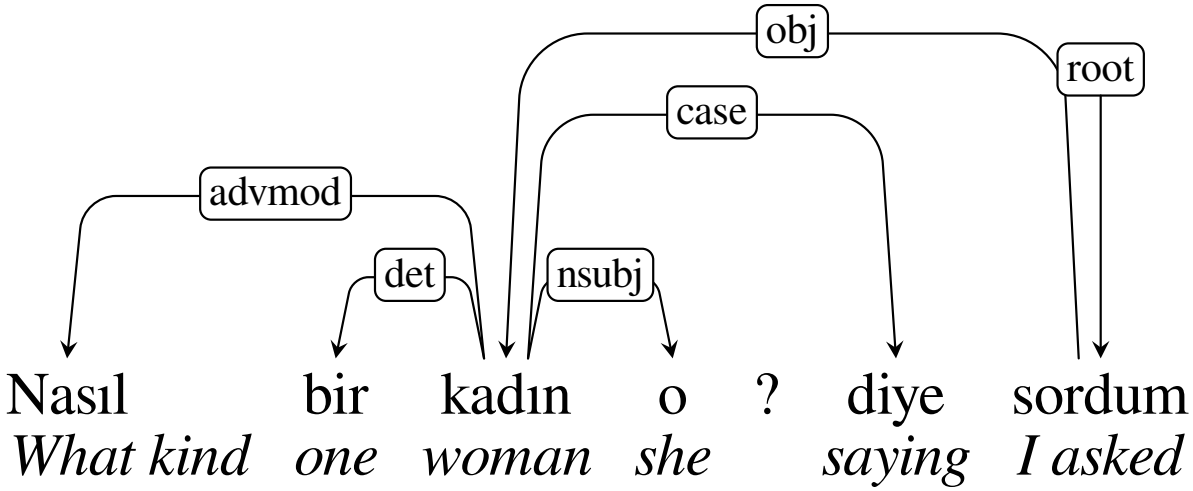
E.g., treatment of denominal adjectives: productive or not?

- Turkish: dağlı = dağ `NOUN` + lı `ADP`
- Kazakh: таулы = таулы `ADJ`, also тау `NOUN` + лы `ADP`
- Uyghur: تاغلىق `NOUN`

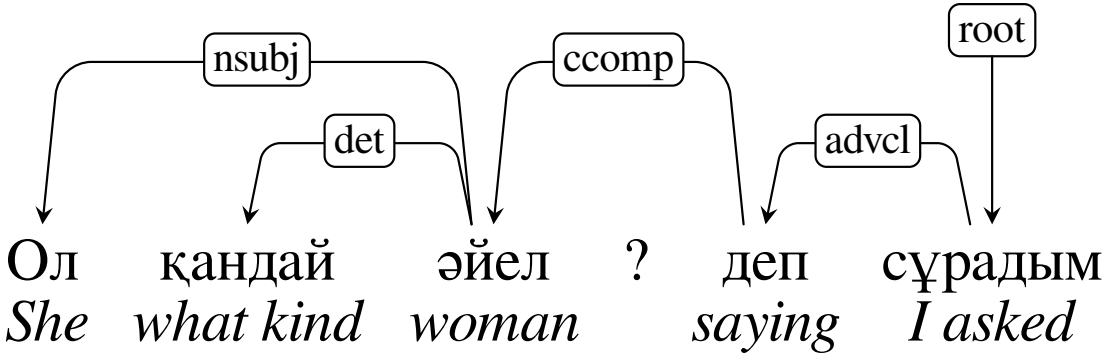
Turkic languages in UD

Differences between the treebanks: dependency relations

Turkish: diye as `case` dependent of verbal object



Kazakh: деп as head of `advcl` with `ccomp` dependency



Turkic languages in UD

Differences between the treebanks: language-specific tags

Relation	Comments	Kazakh	Turkish	Uyghur
`acl:poss`	Adnominal modification with possessive	✓	–	–
`acl:relcl`	Adnominal modification with verbal adjective	✓	–	–
`advmod:emph`	Adverbial emphasiser (mostly -dA)	–	✓	✓
`aux:q`	Question word, -mI	–	✓	–
`compound:lvc`	Light verb	✓	✓	✓
`compound:redup`	Reduplication compound	–	✓	✓
`flat:name`	Proper name	✓	–	–
`iobj:caus`	Causee	✓	–	–
`nmod:abl`	Oblique in the ablative	*	*	✓
`nmod:cau`	Causee	*	*	✓
`nmod:clas`	Noun-noun compound	*	*	✓
`nmod:comp`	Nominal modifier [mostly ablative]	–	–	✓
`nmod:poss`	Genitive possessive modifier	✓	✓	✓
`nmod:tmod`	Time modifier	–	–	✓
`obl:own`	Owner in -DA	✓	–	–

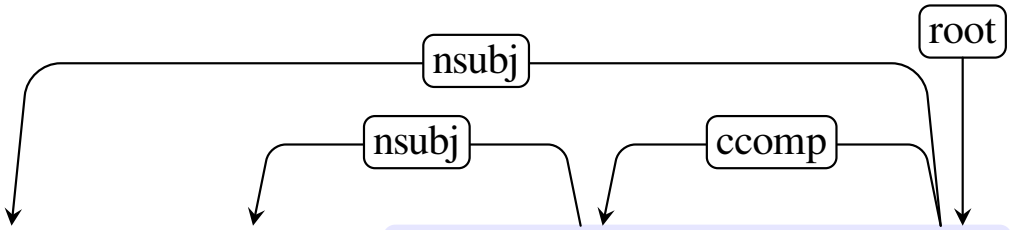
Parsing performance

Parsing performance in the CoNLL shared task.

Language	Train	Dev	Winning team (LAS)	UAS	LAS	CLAS
Kazakh	0	529	Dumitrescu et al. (2017)	45.72	29.22	25.14
Turkish	38 082	10 011	Dozat et al. (2017)	69.62	62.79	60.01
Turkish-PUD	38 082	10 011	Björkelund et al. (2017)	59.35	38.22	32.32
Uyghur	0	1662	Björkelund et al. (2017)	60.57	43.51	34.07

Open questions

Tokenisation



	Örnek	bizim	yazdıklarımızdan	-dı
Gloss	*example*	*we*-GEN	*wrote*-PART.1PL	*was*-3SG
POS	NOUN	PRON	VERB	VERB
Lemma	örnek	biz	yaz	i-
Number	Plur	Plur	Plur	Sing
Case	Nom	Gen	Abl	-
Person	3	1	3	3
Number[psor]	-	-	Plur	-
Person[psor]	-	-	1	-
VerbForm	-	-	Part	-
Tense	-	-	Past	Past

Open questions

Core and oblique

In UD:

- `obj` is the most core element after `subj`;
- `iobj` is the most core element after `obj`;
- oblique (`obl`) is a non-core `obj`.

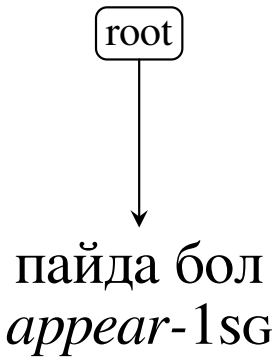
In Turkic languages:

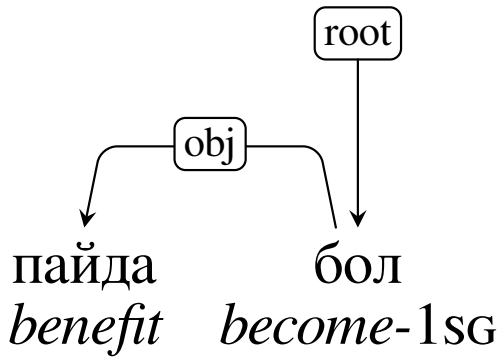
- nothing is mandatory not even subject;
- possible test: passive/causative case promotion/demotion;

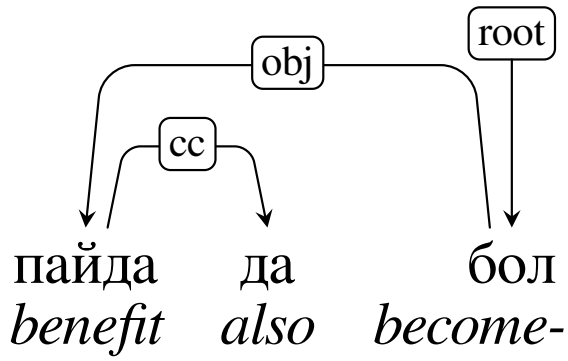
Open questions

Complex predicates

Non-verbal + Verbal



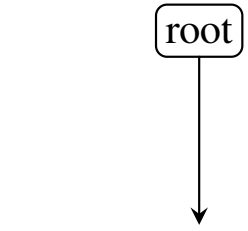




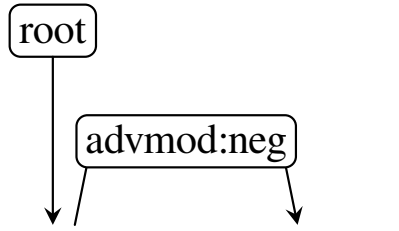
Open questions

Complex predicates

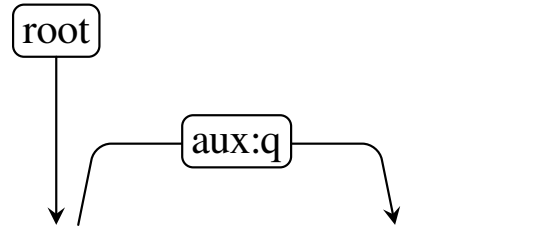
Verbal + Non-verbal



(a) Current analysis of Kazakh multi-token negation



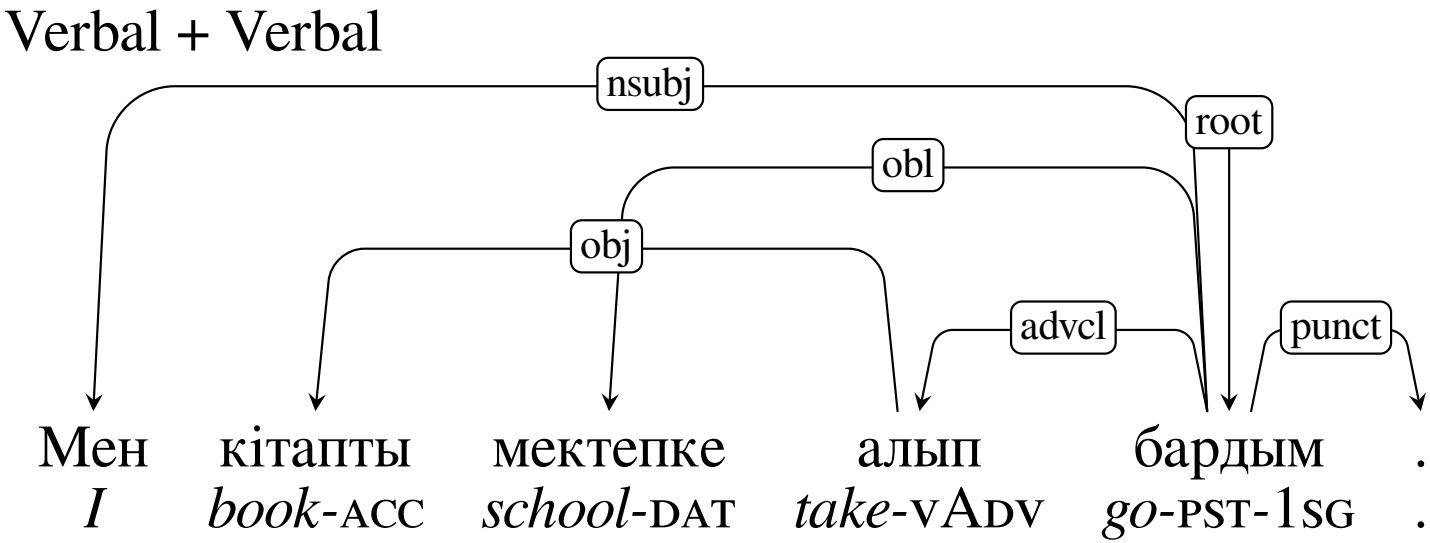
(b) Alternative proposal



(c) Turkish multi-token question word

Open questions

Complex predicates



Open questions

Multiple derivation

Multiple values for voice (a) and aspect (b):

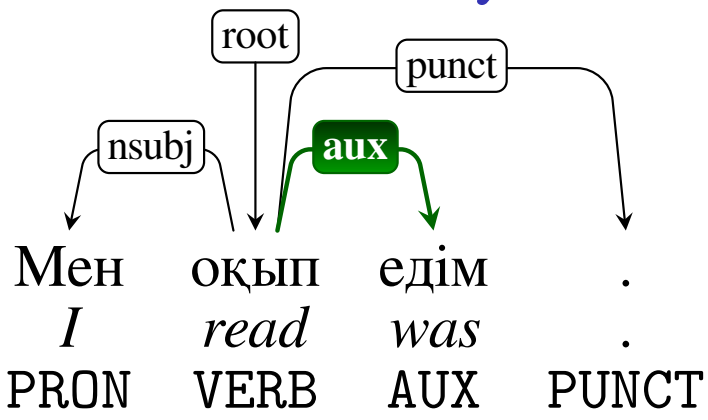
a. *bekle -t -il -iyor*
wait CAUS PASS PROG
'being stalled (=caused to wait)'

b. *oku -yuver -iyor*
read RAPID PROG
'he/she is reading quickly'

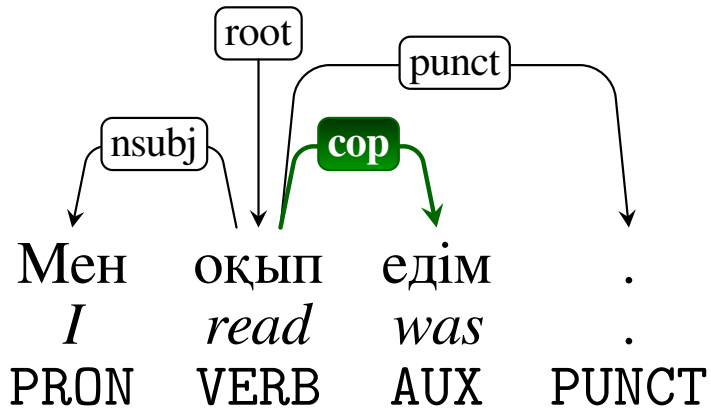
Open questions

Use of copulas with non-finite verb forms

Converb + Auxiliary?



Converb + Copula?



Concluding remarks

- Existing corpora have some differences in annotation
- Mostly due to conversion from different grammatical traditions
- Better coordination among Turkic annotators needed
- UD is an effective standard for all Turkic languages

Acknowledgements

PƏXMƏT!

Also:

- UD community for thoughtful discussion and input on a range of issues discussed in this paper
- Deniz Uysal and Tolgonay Kubatova for help with native speaker judgements.
- Jonathan Washington (from the presenter) for doing most of these great slides.