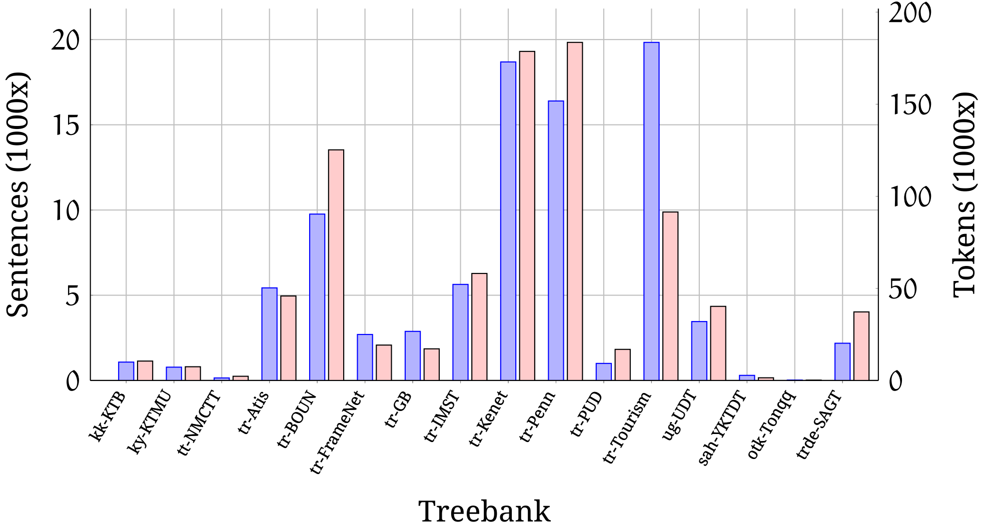
Turkic UD treebanks

Overview, common issues

Büşra Marşan, Furkan Akkurt, Çağrı Çöltekin

September 8, 2023

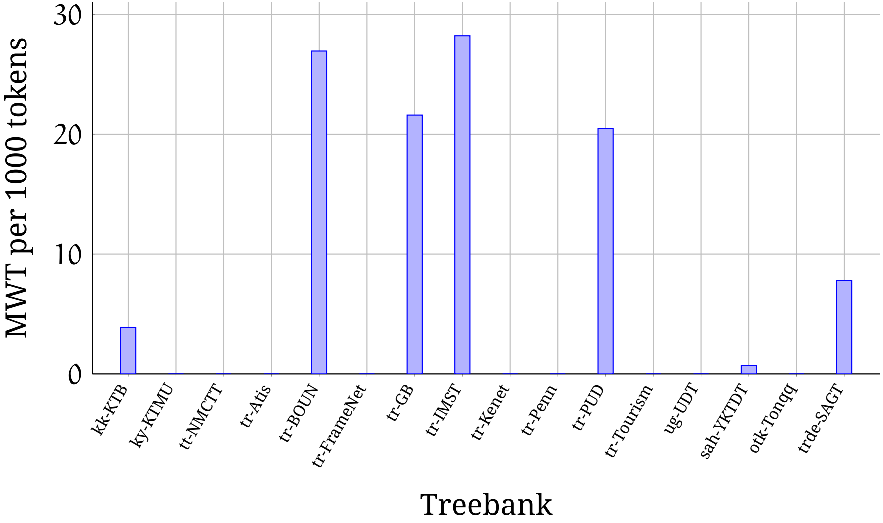
Current treebanks (as of UD 1.12)



Discussion points / issues

- segmentation/MWE
    - compounds, two-part words
    - -ki
- feature specification
- copula, copula as auxiliary
- oblique/object distinction
- question particle
- converb, non-finite verb forms
- ‘periphrastic’ negative finite verb forms (kaz/kir: barğan joqsuŋ, barğan emessiŋ, tat: barğanıŋ yuq)
- code-switching
- cross-lingual/historical consistency
- semantic representation
- root in parataxis, compound sentences
- adpositions

Multi-word token distribution

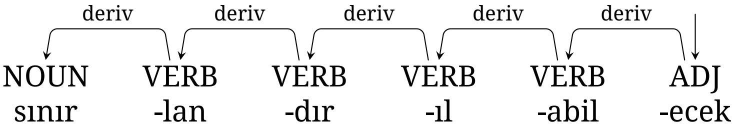


What is segmented (currently)?

- Copular markers *küçük-sün*, *var-dı*, *siyasetçi-ydi*, *tutuyorlar-dı*, *olmayacak-tır*, *қуанышты-мын* all (BOUN, GB, IMST, PUD, SAGT, KTB)
    - Some treebanks do not split copular affixes attached to verbal forms (e.g., *tutuyorlar-dı*, *olmayacak-tır*)
- -ki *yüzeyinde-ki кім-дікі* (BOUN, GB, IMST, PUD, SAGT, KTB)
- -li *(sarı) saç-lı (бір) палата-лы* (BOUN, GB, SAGT, KTB)
- -siz *(renkli) cam-sız* (BOUN, GB, SAGT)
- -lik *(bin) lira-lık* (BOUN, GB, SAGT)
- -(y)ici *can al-ıcı* (IMST)
- -ce *(yöre) halkı-nca* (GB)

Why do we split (written) words?

- The ‘syntactic words’ are multiple nodes in a parse tree
  isn’t = *is* + *not*
- History in Turkish dependency annotation: *inflectional groups*



deriv deriv deriv deriv deriv

NOUN VERB VERB VERB VERB ADJ

sınır -lan -dır -ıl -abil -ecek

- Current practice is more conservative
- Other extreme: no word segmentation at all
- Note: currently there is an ongoing discussion on ‘word’ in UD

Need for sub-word units: an example with suffix *-ki*

Yan	odadakiler	uyuyorlar
Side	*room-in-the-ones*	*sleep*

‘The ones in the next room are sleeping’

- *oda* is singular, *odadakiler* (people in the room) are plural
- *yan* modifies only *oda*, not the people
- The issue is not present in adjectival uses of *-ki*
- *-ki* may repeat (*odada**ki**lerin**ki***

Need for sub-word units: an example with suffix *-ki*

	Yan	odadakiler	uyuyorlar
	Side	*room-in-the-ones*	*sleep*
Lemma:	yan	oda	sleep

‘The ones in the next room are sleeping’

- *oda* is singular, *odadakiler* (people in the room) are plural
- *yan* modifies only *oda*, not the people
- The issue is not present in adjectival uses of *-ki*
- *-ki* may repeat (*odada**ki**lerin**ki***

Need for sub-word units: an example with suffix *-ki*

	Yan	odadakiler	uyuyorlar
	Side	*room-in-the-ones*	*sleep*
Lemma:	yan	oda	sleep
POS:	ADJ	NOUN	VERB

‘The ones in the next room are sleeping’

- *oda* is singular, *odadakiler* (people in the room) are plural
- *yan* modifies only *oda*, not the people
- The issue is not present in adjectival uses of *-ki*
- *-ki* may repeat (*odada**ki**lerin**ki***

Need for sub-word units: an example with suffix *-ki*

	Yan	odadakiler	uyuyorlar
	Side	*room-in-the-ones*	*sleep*
Lemma:	yan	oda	sleep
POS:	ADJ	NOUN	VERB
Number:	-	plural	plural
Person:	-	3	3

‘The ones in the next room are sleeping’

- *oda* is singular, *odadakiler* (people in the room) are plural
- *yan* modifies only *oda*, not the people
- The issue is not present in adjectival uses of *-ki*
- *-ki* may repeat (*odada**ki**lerin**ki***

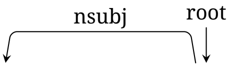
Need for sub-word units: an example with suffix *-ki*

			root ↓
	Yan	odadakiler	uyuyorlar
	Side	*room-in-the-ones*	*sleep*
Lemma:	yan	oda	sleep
POS:	ADJ	NOUN	VERB
Number:	-	plural	plural
Person:	-	3	3

‘The ones in the next room are sleeping’

- *oda* is singular, *odadakiler* (people in the room) are plural
- *yan* modifies only *oda*, not the people
- The issue is not present in adjectival uses of *-ki*
- *-ki* may repeat (*odada**ki**lerin**ki***

Need for sub-word units: an example with suffix *-ki*

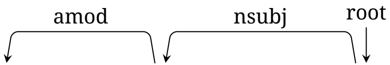


	Yan	odadakiler	uyuyorlar
	Side	*room-in-the-ones*	*sleep*
Lemma:	yan	oda	sleep
POS:	ADJ	NOUN	VERB
Number:	-	plural	plural
Person:	-	3	3

‘The ones in the next room are sleeping’

- *oda* is singular, *odadakiler* (people in the room) are plural
- *yan* modifies only *oda*, not the people
- The issue is not present in adjectival uses of *-ki*
- *-ki* may repeat (*odada**ki**lerin**ki***

Need for sub-word units: an example with suffix *-ki*



	Yan	odadakiler	uyuyorlar
	Side	*room-in-the-ones*	*sleep*
Lemma:	yan	oda	sleep
POS:	ADJ	NOUN	VERB
Number:	-	plural	plural
Person:	-	3	3

‘The ones in the next room are sleeping’

- *oda* is singular, *odadakiler* (people in the room) are plural
- *yan* modifies only *oda*, not the people
- The issue is not present in adjectival uses of *-ki*
- *-ki* may repeat (*odada**ki**lerin**ki***

Words with spaces

- UD has three relations for MWE: `fixed`, `flat` and `compound`
- The constructions of interest here include *light verb constructions clitics*, *compounds*
- Currently, KTB has some word forms with spaces (*естіген жоқ екен*)
- A possible direction for consistency may be unifying the forms that are written differently in some languages (e.g., question particle)

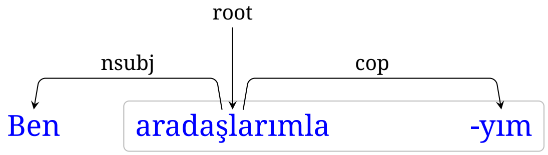
Morphological feature specification

- Multiple features (generally verbal) on the same verb:
  - *soğu-**t-ul**-ur* ‘ (made to) be cooled’ `Voice=CauPass`
- Some features may repeat (currently no solution):
  - *oku-**ya**-ma-**yabil**-ir-im* ‘I **may** not be **able to** read’
  - *oku-**n-ul**-ma-z* ‘One may/can not read’ (zero person / impersonal passive)
  - *oku-**t-tur**-du* ‘s/he caused/made someone else to cause/make someone to read’
- If not segmented, features for affixes like *-lI*, *-sIz*:
  - *araba-m-**sız*** ‘**without** my car’
- Lexicalized/productive use of some affixes (Like *-lI*, *-sIz* above, but also reflexive, reciprocal):
  - *bul-uş-* ‘to meet (to find each other)’ – *öpü-ş-* ‘to kiss (each other)’ – *selamla-ş-* ‘to greet each other’
- TAME assignment is currently (very) inconsistent
- Nominal inflections on adjectivals

Copular constructions

- Copular suffix is segmented inconsistently
  - No split
  - Split all copular suffixes
  - Split only copular suffixes attached to nominals
- Segmentation requires null-tokens when copula is not realized (third-person, singular)

Null copula: an example



root

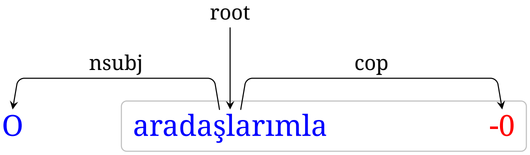
nsubj

cop

Ben	aradaşlarımla	-yım
I	*friend.PL.P1S.INS*	*COP.PRES.1SG*

‘I am with my friends’

Null copula: an example



root

nsubj

cop

O aradaşlarımla -0

He/she *friend.PL.P1S.INS* *COP.PRES.3.SG*

‘He/she is with my friends’

Core vs. non-core

- Argument–adjunct distinction is useful for some applications
- UD makes distinctions between core (object) and non-core (oblique) modifiers of predicates
- UD guidelines suggests case marking as a guide for determining core/non-core
- A possible way forward is tests for ‘coreness’

Object cases in current treebanks

KTB	Acc, Dat, Nom
KTMU	Abl, Acc, Dat, Dat,Gen, Gen, Ins, Nom
NMCTT	Acc, Nom
Atis	Abl, Acc, Dat, Ins, Nom
BOUN	Abl, Acc, Dat, Gen, Ins, Loc, Nom
FrameNet	Abl, Acc, Dat, Gen, Ins, Loc, Nom
GB	Abl, Acc, Nom
SAGT	Acc, Dat, Ins, Nom
IMST	Abl, Acc, Dat, Equ, Gen, Ins, Loc, Nom
Kenet	Abl, Acc, Dat, Gen, Ins, Loc, Nom
Penn	Abl, Acc, Dat, Gen, Ins, Loc, Nom
PUD	Abl, Acc, Dat, Gen, Ins, Loc, Nom
Tourism	Abl, Acc, Dat, Gen, Ins, Loc, Nom
UDT	Abl, Acc, Dat, Loc, Nom
YKTDT	Abl, Acc, Dat, Ins, Nom, Par

Indirect object cases in current treebanks

KTB	Abl, Acc, Dat
BOUN	Abl, Acc, Dat, Ins, Nom
FrameNet	Dat
SAGT	Acc, Dat
IMST	Abl, Acc, Dat, Gen, Ins, Loc, Nom
Kenet	Abl, Acc, Dat, Gen, Nom
Penn	Dat, Nom
PUD	Dat
UDT	Dat
YKTDT	Dat, Ins

Question particle

- The writing standards for the question particle differs among Turkic languages
- When considered as a separate token, there is no clear way to annotate question particle in UD
- Most treebanks use `AUX` tag, and `aux` relation, since in some cases (but not all) TAME markers may follow the question particle

Other points from participants

- converb, non-finite verb forms
- ‘periphrastic’ negative finite verb forms
- code-switching
- cross-lingual/historical consistency
- semantic representation
- root in parataxis, compound sentences
- adpositions

Tense

	Fut	Fut,Past	NearPast	Past	PastPerf	PastResultI	Pqp	Pres
KTB	✓			✓				✓
KTMU	✓			✓				✓
Tonqq								
NMCTT	✓			✓				✓
Atis	✓			✓				✓
BOUN	✓			✓				✓
FrameNet	✓			✓				✓
GB	✓	✓		✓			✓	✓
SAGT	✓			✓			✓	✓
IMST	✓			✓			✓	✓
Kenet	✓			✓				✓
Penn	✓			✓				✓
PUD	✓			✓			✓	✓
Tourism	✓			✓				✓
UDT				✓				✓
YKTDT	✓		✓	✓	✓	✓		✓

Aspect

	Dur	Hab	Imp	Iter	Perf	Prog	Prosp	Rapid
KTB		✓	✓		✓			
KTMU					✓	✓		
Tonqq								
NMCTT				✓	✓	✓		
Atis		✓			✓	✓		
BOUN		✓	✓		✓	✓	✓	
FrameNet		✓			✓	✓		
GB	✓	✓	✓		✓	✓	✓	✓
SAGT		✓	✓		✓	✓	✓	
IMST	✓	✓			✓	✓	✓	✓
Kenet		✓			✓	✓		✓
Penn		✓			✓	✓		
PUD		✓			✓	✓	✓	
Tourism		✓			✓	✓		
UDT		✓			✓			
YKTDT								

Mood

	Cnd	CndGen	CndGenPot	CndPot	Des	DesPot	Gen	GenNec	GenNecPot	GenPot	GenPotPot	Imp	Ind	Int	Irr	Nec	NecPot	Opt	Pot	PotPot	Prs	Sub
KTB	✓				✓							✓	✓					✓	✓			
KTMU	✓											✓	✓						✓			
Tonqq																						
NMCTT	✓											✓	✓		✓							
Atis	✓	✓		✓			✓			✓		✓	✓			✓		✓	✓			
BOUN	✓				✓	✓	✓	✓		✓		✓	✓			✓		✓	✓			
FrameNet	✓	✓	✓		✓		✓	✓		✓		✓	✓			✓		✓	✓			
GB	✓	✓	✓	✓			✓	✓		✓	✓	✓	✓			✓		✓	✓	✓		
SAGT	✓				✓		✓			✓		✓	✓			✓		✓	✓		✓	✓
IMST	✓			✓	✓	✓	✓	✓	✓	✓		✓	✓			✓	✓	✓	✓		✓	
Kenet	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	✓			✓		✓	✓			
Penn	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	✓			✓		✓	✓			
PUD	✓	✓		✓			✓	✓		✓		✓	✓			✓		✓	✓			
Tourism	✓	✓	✓		✓		✓			✓		✓	✓			✓		✓	✓			
UDT	✓											✓	✓	✓								
YKTDT												✓										

Evident

	Fh	Nfh
KTB	✓	
KTMU	✓	
Tonqq		
NMCTT		
Atis		
BOUN	✓	✓
FrameNet		
GB	✓	✓
SAGT	✓	✓
IMST		✓
Kenet		
Penn		
PUD		✓
Tourism		
UDT		
YKTDT		✓

blank