A Collaborative Web Tool for Linguistic Annotation

Salih Furkan Akkurt

Boğaziçi University - CMPE 492 Project - Advisor: Suzan Uskudarli

DEPARTMENT OF COMPUTER ENGINEERING

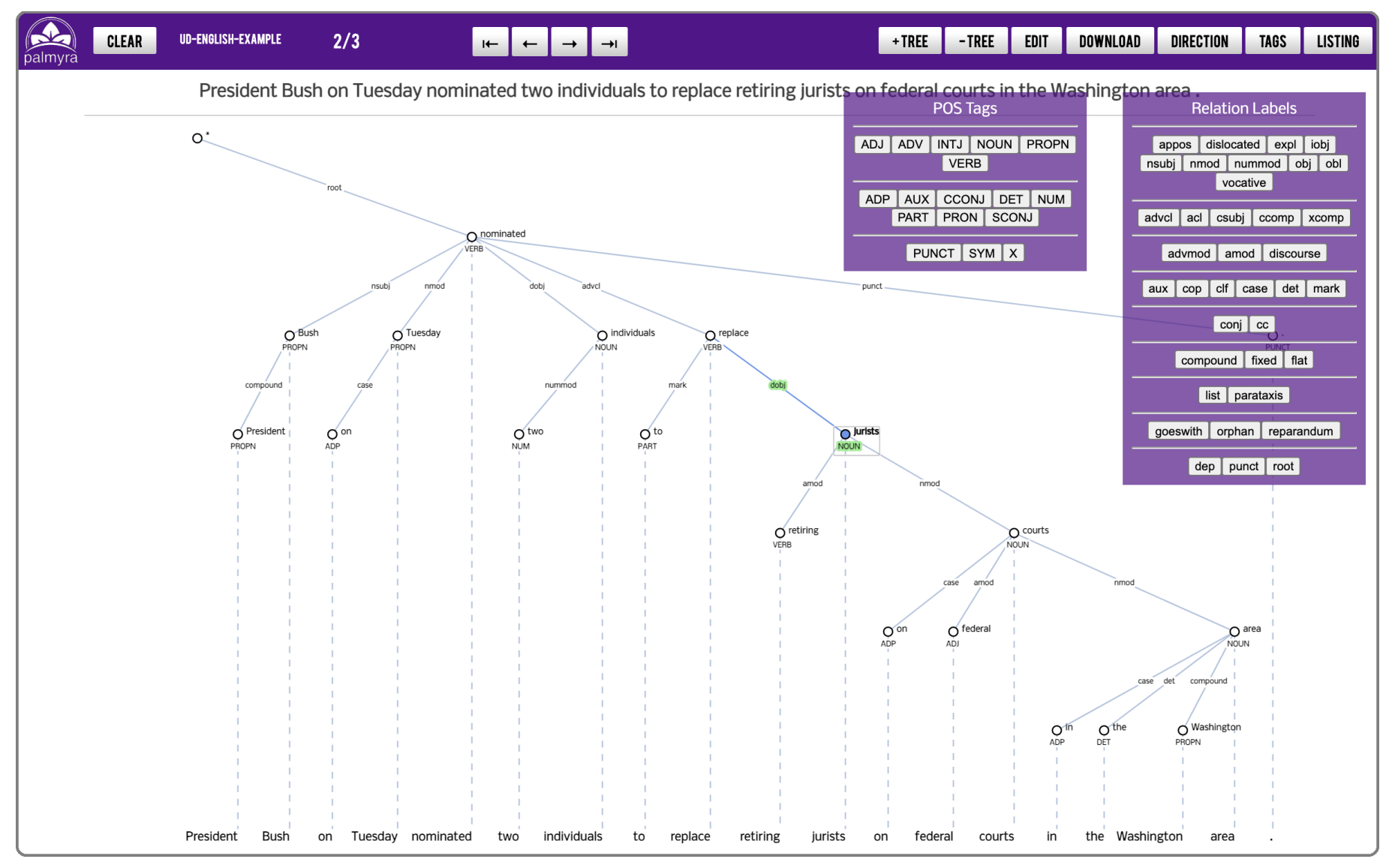
Problem Statement

To design and develop a web-based tool that supports grammatical annotations of morphologically rich languages (MRLs), centers around annotator's user experience and is collaborative, open-source and accessible.

Motivation

Large annotated datasets are crucial to NLP solutions. Creation of quality treebanks takes time and effort. Existing tools are not sufficient for annotating MRLs. Also, treebanks are annotated with multiple annotators, making collaborative tools necessary.

Related Work



Drag-drop interface for annotation [1]

Boğaziçi University Annotation Tool (BoAT): An annotation tool released in 2019 by the CMPE Department of Boğaziçi University. It was designed to annotate MRLs and has been used to annotate the BOUN Treebank, which has 9,761 sentences and is published on UD[2]. It's a standalone application.

Requirements Elicitation & Management

Based on the experience with the BOUN Treebank and numerous meetings with an annotator with linguistic background, we decided to redesign BoAT with these 3 main themes: (1) Addressing annotator requests and improving existing functionality; (2) Extending features with a collaborative experience; (3) New architecture that supports persistence, accessibility and maintainability.

Sentence Example for Agglutinative vs Analytical Languages

MRLs require extensive and complex annotations, illustrated with the following two examples:

English sentence: Most people would find airport wi-fi more useful than being able to send emails *on a plane*.[3]

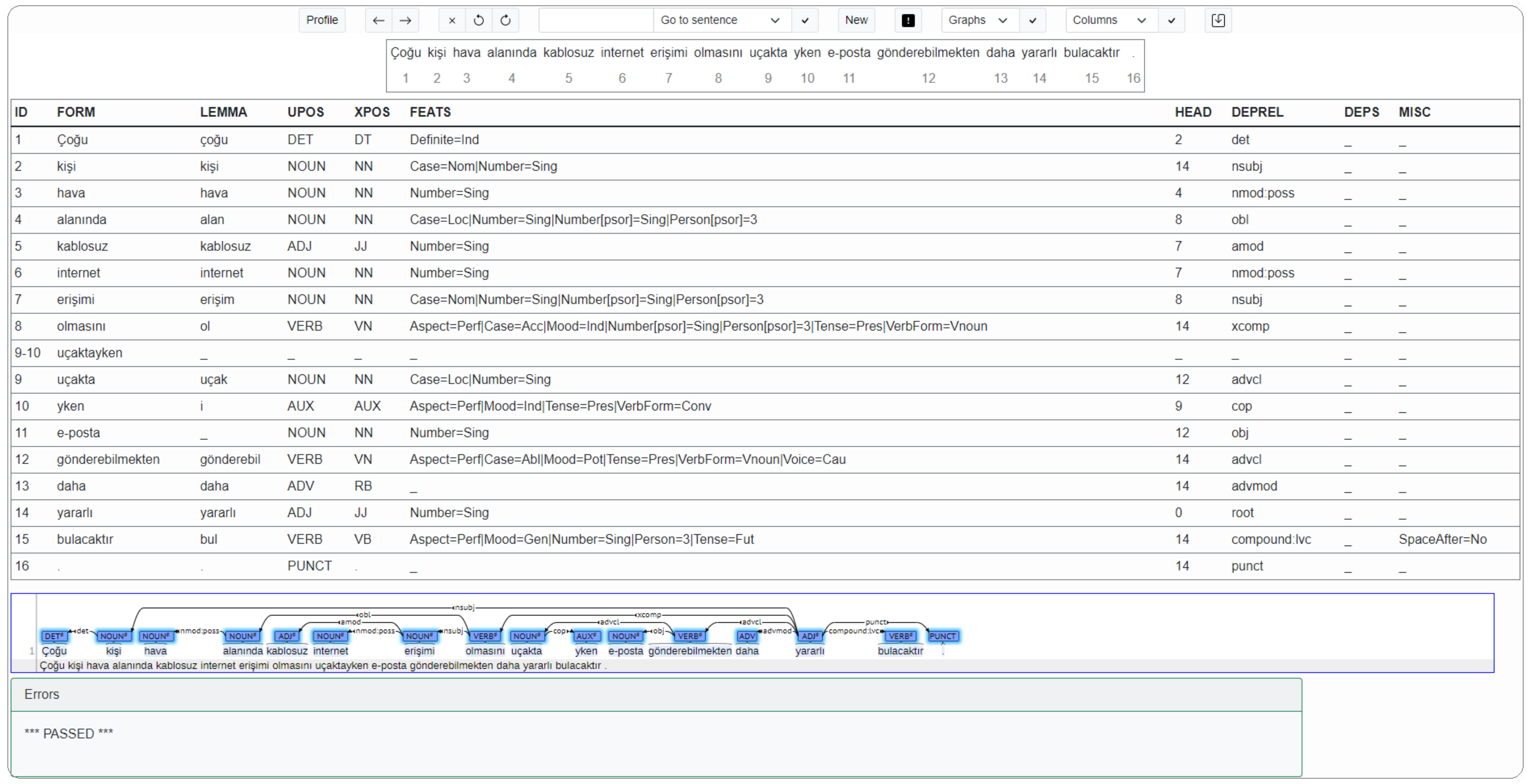
ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL
15	on	on	ADP	IN	_	18	case
16	a	a	DET	DT	Definite=Ind|PronType=Art	18	det
17	plane	plane	NOUN	NN	Number=Sing	14	obl

Turkish sentence: Çoğu kişi hava alanında kablosuz internet erişimi olmasını *uçaktayken* e-posta gönderebilmekten daha yararlı bulacaktır.[4]

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL
9-10	uçaktayken	_	_	_	_	_	_
9	uçakta	uçak	NOUN	NN	Case=Loc|Number=Sing	12	advcl
10	yken	i	AUX	AUX	Aspect=Perf|Mood=Ind|Tense=Pres|VerbForm=Conv	9	cop

Solution

We implemented a tool for MRLs like Turkish.



Çoğu kişi hava alanında kablosuz internet erişimi olmasını uçakta yken e-posta gönderebilmekten daha yararlı bulacaktır .
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

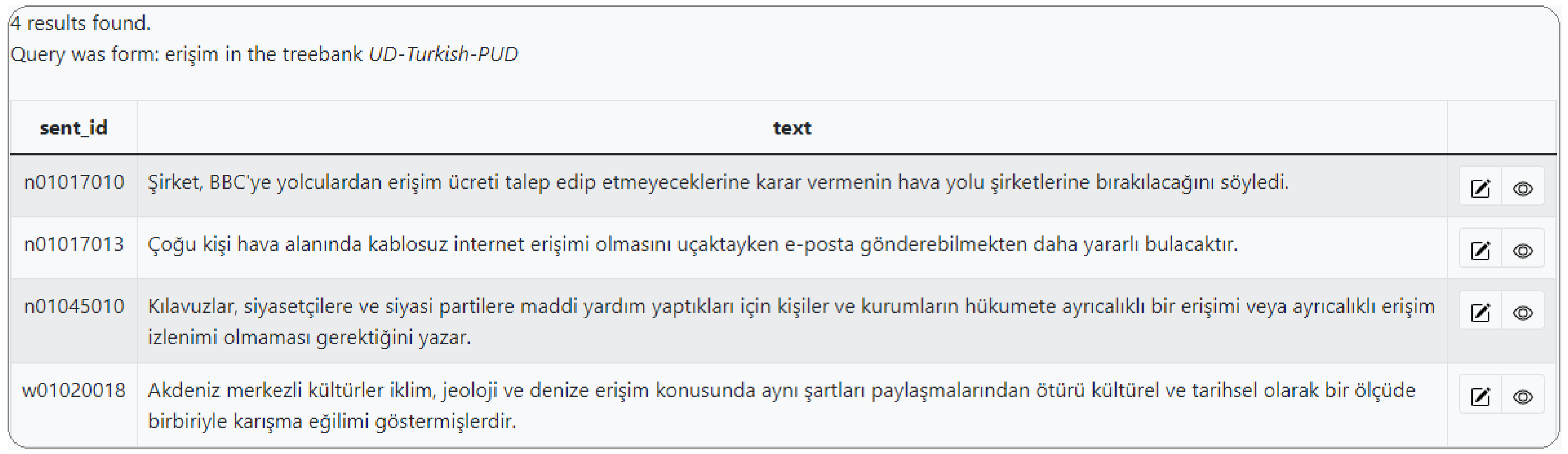
ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Çoğu	çoğu	DET	DT	Definite=Ind	2	det	_	_
2	kişi	kişi	NOUN	NN	Case=Nom|Number=Sing	14	nsubj	_	_
3	hava	hava	NOUN	NN	Number=Sing	4	nmod:poss	_	_
4	alanında	alan	NOUN	NN	Case=Loc|Number=Sing|Number[psor]=Sing|Person[psor]=3	8	obl	_	_
5	kablosuz	kablosuz	ADJ	JJ	Number=Sing	7	amod	_	_
6	internet	internet	NOUN	NN	Number=Sing	7	nmod:poss	_	_
7	erişimi	erişim	NOUN	NN	Case=Nom|Number=Sing|Number[psor]=Sing|Person[psor]=3	8	nsubj	_	_
8	olmasını	ol	VERB	VN	Aspect=Perf|Case=Acc|Mood=Ind|Number[psor]=Sing|Person[psor]=3|Tense=Pres|VerbForm=Vnoun	14	xcomp	_	_
9-10	uçaktayken	_	_	_	_	_	_	_	_
9	uçakta	uçak	NOUN	NN	Case=Loc|Number=Sing	12	advcl	_	_
10	yken	i	AUX	AUX	Aspect=Perf|Mood=Ind|Tense=Pres|VerbForm=Conv	9	cop	_	_
11	e-posta	_	NOUN	NN	Number=Sing	12	obj	_	_
12	gönderebilmekten	gönderebil	VERB	VN	Aspect=Perf|Case=Abl|Mood=Pot|Tense=Pres|VerbForm=Vnoun|Voice=Cau	14	advcl	_	_
13	daha	daha	ADV	RB	_	14	advmod	_	_
14	yararlı	yararlı	ADJ	JJ	Number=Sing	0	root	_	_
15	bulacaktır	bul	VERB	VB	Aspect=Perf|Mood=Gen|Number=Sing|Person=3|Tense=Fut	14	compound:lvc	_	SpaceAfter=No
16	.	.	PUNCT	.	_	14	punct	_	_

1 Çoğu kişi hava alanında kablosuz internet erişimi olmasını uçaktayken e-posta gönderebilmekten daha yararlı bulacaktır .

Errors

```
*** PASSED ***
```

Annotation of "Çoğu kişi hava alanında kablosuz internet erişimi olmasını uçaktayken e-posta gönderebilmekten daha yararlı bulacaktır.", using BoAT v2



4 results found.
Query was form: erişim in the treebank *UD-Turkish-PUD*

sent_id	text	
n01017010	Şirket, BBC'ye yolculardan erişim ücreti talep edip etmeyeceklerine karar vermenin hava yolu şirketlerine bırakılacağını söyledi.	
n01017013	Çoğu kişi hava alanında kablosuz internet erişimi olmasını uçaktayken e-posta gönderebilmekten daha yararlı bulacaktır.	
n01045010	Kılavuzlar, siyasetçilere ve siyasi partilere maddi yardım yaptıkları için kişiler ve kurumların hükumete ayrıcalıklı bir erişimi veya ayrıcalıklı erişim izlenimi olmaması gerektiğini yazar.	
w01020018	Akdeniz merkezli kültürler iklim, jeoloji ve denize erişim konusunda aynı şartları paylaşmalarından ötürü kültürel ve tarihsel olarak bir ölçüde birbiriyle karışma eğilimi göstermişlerdir.	

Results from searching the "form" field as "erişim" for the treebank *UD-Turkish-PUD*

Result

Presented at The International Conference on Agglutinative Language Technologies as a Challenge of Natural Language Processing (ALTNLP) in Jun 2022 [5], which was well-received, resulting in several downloads.

Currently, acceptance testing is being performed with annotators with linguistic background. Preliminary results are promising.

A dockerized version is available on GitHub [6].

Future Directions

- Endangered languages in the Linguistics Department of Boğaziçi University.
- Uzbek: Offer came to collaborate during ALTNLP 2022.
- Automated partial annotations.

Acknowledgements

I am grateful to the NLP team of folks from both CMPE and LING departments. A special thanks is due to Büşra Marşan who gave constant feedback and taught me about the annotation process. I am grateful, most of all, to my advisor Ms Uskudarli for spending countless hours with and for me. 🤗

References

[1] `https://universaldependencies.org/tools.html#palmyra`

[2] `https://universaldependencies.org/treebanks/tr_boun/index.html`

[3] `https://universaldependencies.org/treebanks/en_pud/index.html`

[4] `https://universaldependencies.org/treebanks/tr_pud/index.html`

[5] Akkurt, S. F., Marşan B., Uskudarli S. (2022) 'BoAT v2 -A Web-Based Dependency Annotation Tool with Focus on Agglutinative Languages', *The International Conference on Agglutinative Language Technologies as a Challenge of Natural Language Processing*. Slovenia, 7-8 June.

[6] `https://github.com/furkanakkurt1335/boat`

Contact: furkan.akkurt@boun.edu.tr