The more the merrier

A new dependency treebank for Turkish

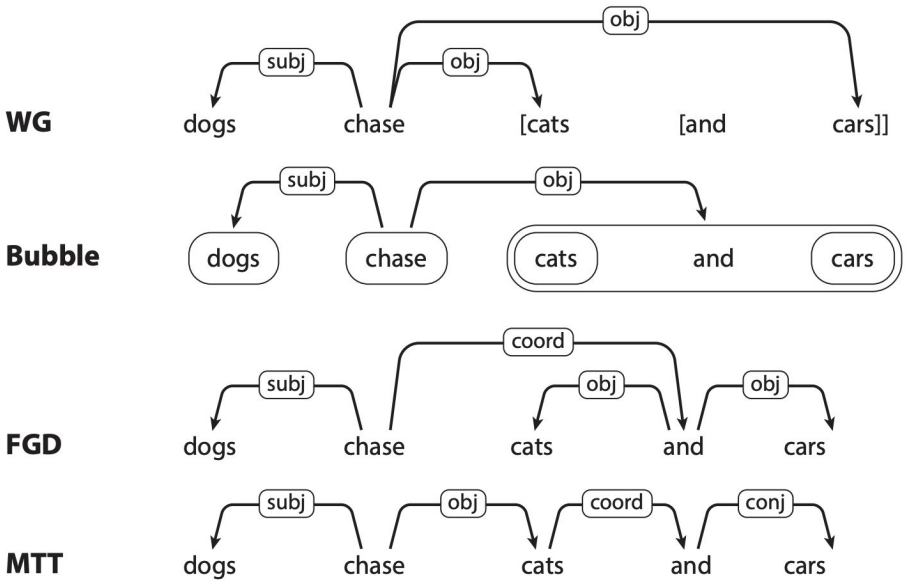
Utku Türk*, Furkan Atmaca*, Şaziye Betül Özateş†, Gözde Berk†, Seyyit Talha Bedir*,

Abdüllatif Köksal†, Balkız Öztürk Başaran*, Tunga Güngör†, Arzucan Özgür†

* Boğaziçi University, Department of Linguistics
† Boğaziçi University, Department of Computer Science

key concepts: dependency grammar

- non-binary structures
- encodes relations between words or word-clusters
- of course, there are bunch of theories
- universal dependencies



Tesnière (1959), Mel’cuk (1988), Milicevic (2006), De Marneffe & Nivre (2019)

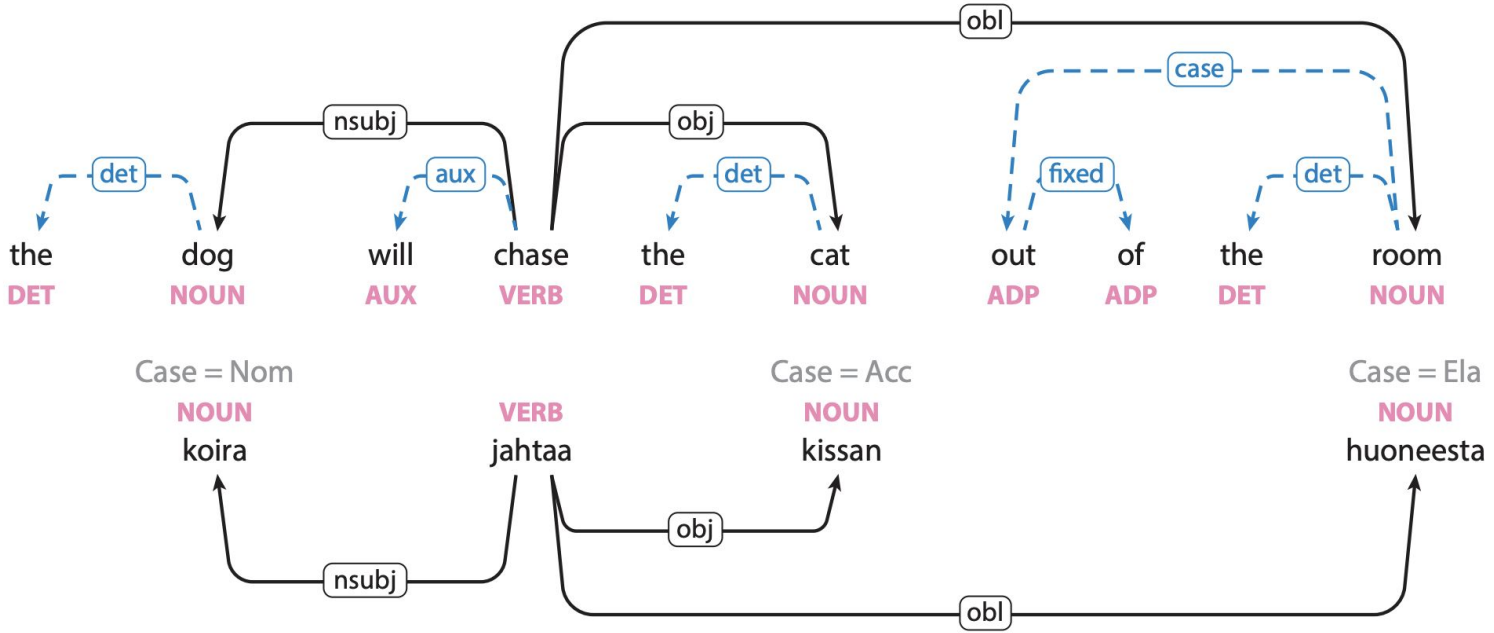
key concepts: universal dependencies

- cross-linguistically consistent
- a limited set of morphological and syntactic tags
- to capture idiosyncrasies and similarities
- open community: < 300 contributors, < 200 treebanks, < 100 languages

---

Nivre et al. (2016), www.universaldependencies.org

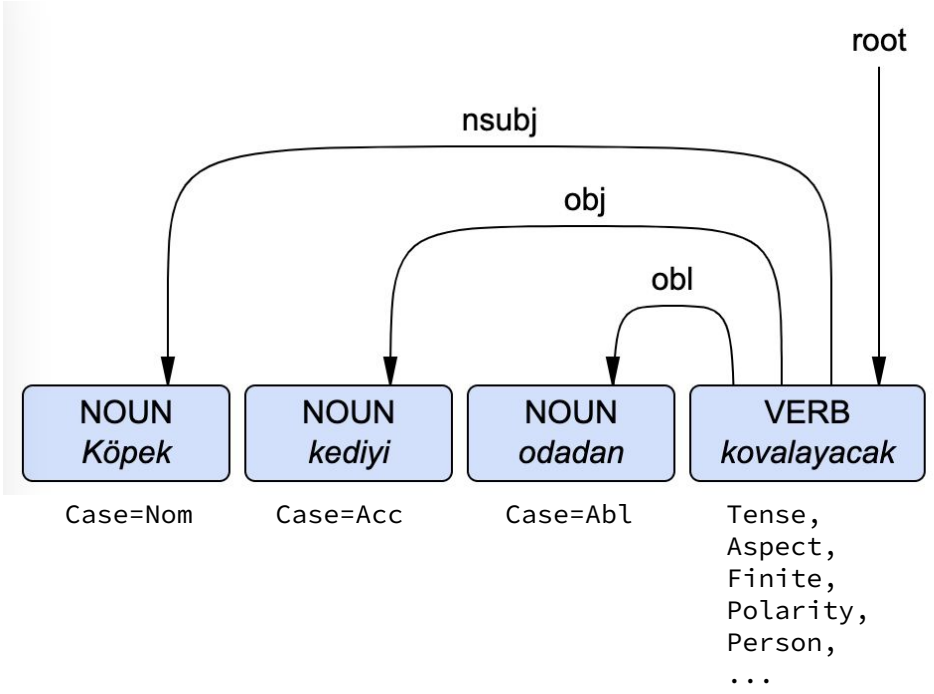
key concepts: universal dependencies



De Marneffe & Nivre (2019)

key concepts: universal dependencies

linguistic details are introduced either using syntactic dependency relations or morphological feature set.



how is Turkish doing?

- not bright until recently

	GB	IMST-UD	PUD	FrameNet	Kenet	Penn	Tourism
N of sentences	2880	5635	1000	2698	18687	9557	19749
N of words	17177	57859	16882	19221	178660	87367	92156
Word per sent.	**5.96**	**10.26**	**16.88**	**7.12**	**9.56**	**9.14**	**4.66**

how is Turkish doing?

- now, it’s alright

	Language	N of words
1.	German	3,753 K
2.	Czech	3,428 K
3.	English	1,880 K
4.	Japanese	1,680 K
…		
14.	**Turkish**	**591 K**

our contribution: BOUN Treebank

	BOUN	GB	IMST-UD	PUD	FrameNet	Kenet	Penn	Tourism
N of sentences	9761	2880	5635	1000	2698	18687	9557	19749
N of words	122383	17177	57859	16882	19221	178660	87367	92156
Word per sent.	12.53	5.96	10.26	16.88	7.12	9.56	9.14	4.66

our contribution: BOUN Treebank

- data gathered from **Turkish National Corpus** (Aksan et al., 2012)

- data distribution:

Genre	N of sentences	N of words
Essays	1953	27557
National Newspaper	1898	29386
Instructional Text	1976	20625
Popular Culture	1962	21263
Biography	1972	23553
Total	**9761**	**122383**

our aims

- proposing our perspective on highly debated topics in UD
  - transparency in embedded clauses,
  - syntactic representation of copular clitics,
  - UD-loyal compound representation,
  - issue of classifiers,
  - what is a core argument?

- providing new data

- new workflow for Turkish UD

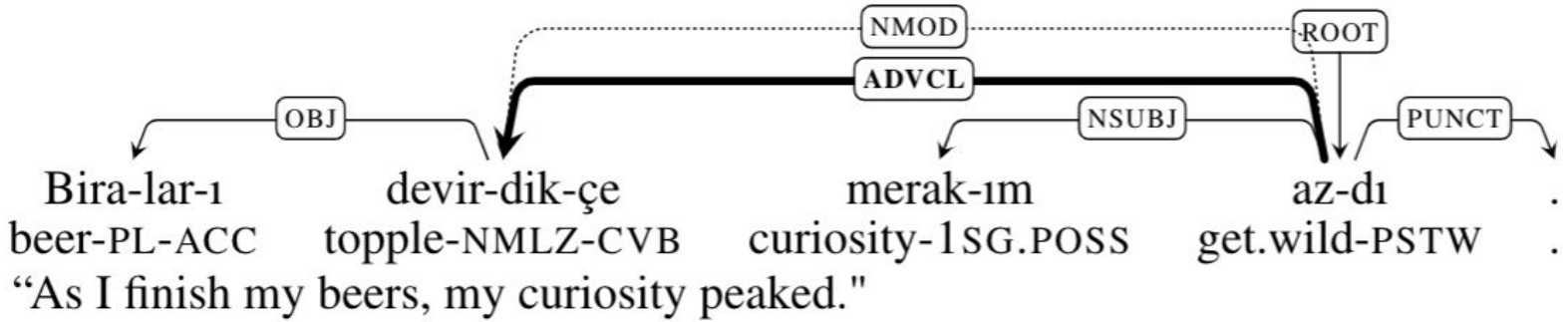
our team & workflow

- 4 linguists and 4 NLP researchers

- discussion on hypothetical sentences
- semi-automatic morphological tagging with Sak et al. (2011)
- individual annotation
- cross-checking and review of annotations
- re-annotation after discussing different applications

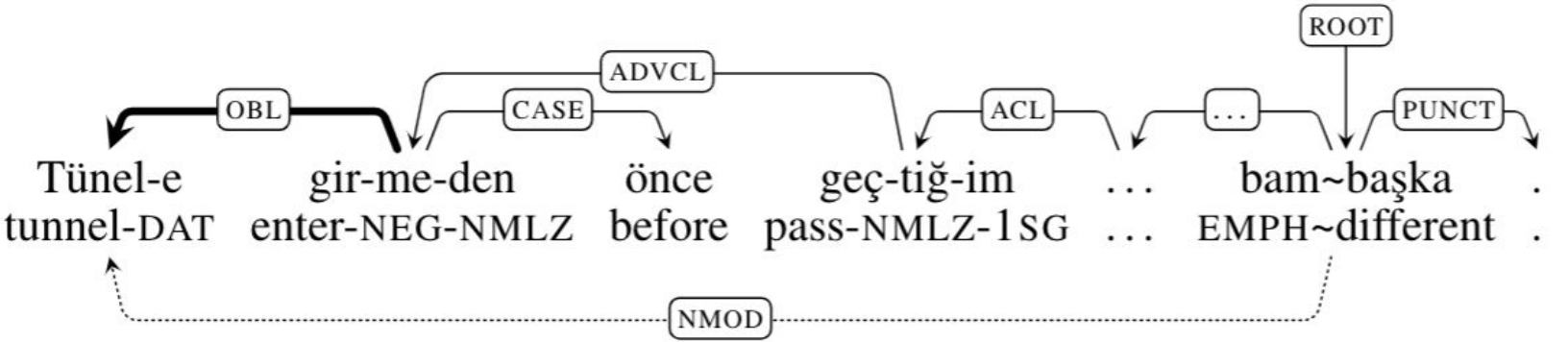
issue #1: embedded sentences

previous annotations does not represent syntactic depth of embedded sentences



issue #1: embedded sentences

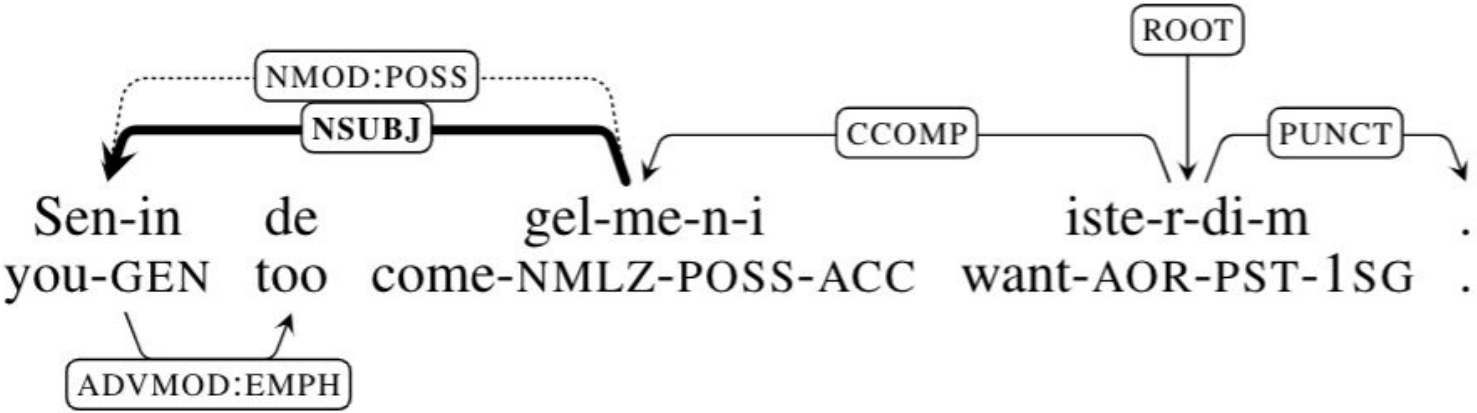
or certain elements are merged to erroneous sites.



'The scenery that I passed before I entered the tunnel was completely different from here.'

issue #1: embedded sentences

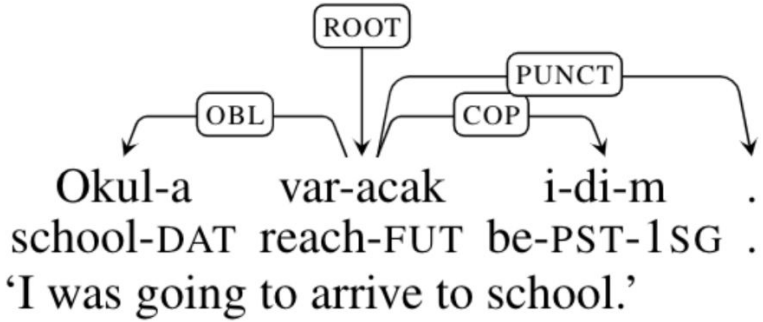
subject or compound?

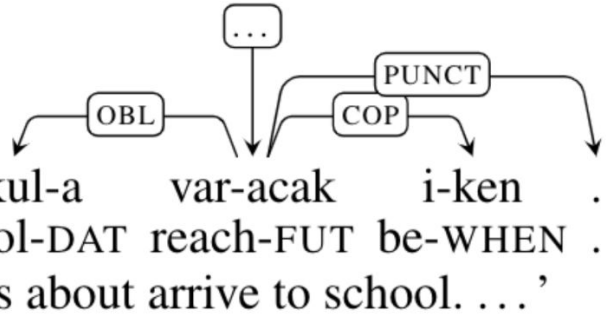


'I would have wanted you to come, as well.'

issue #2: copula

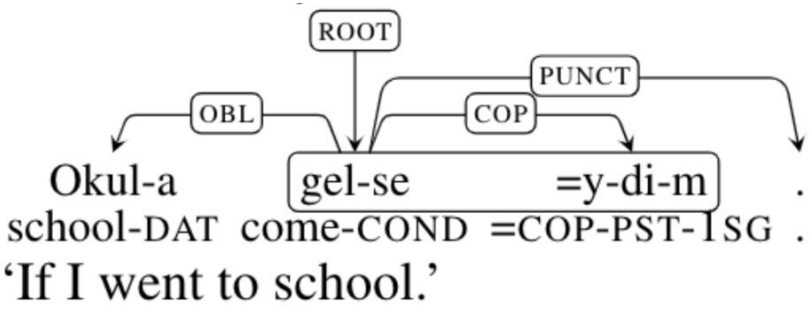
sun is shining bright when the borders of syntactic words are clear

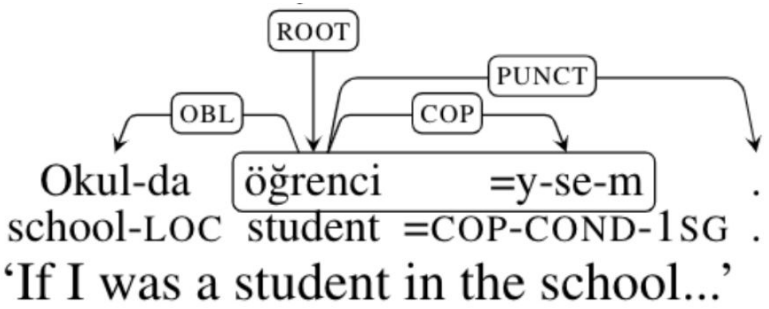




issue #2: copula

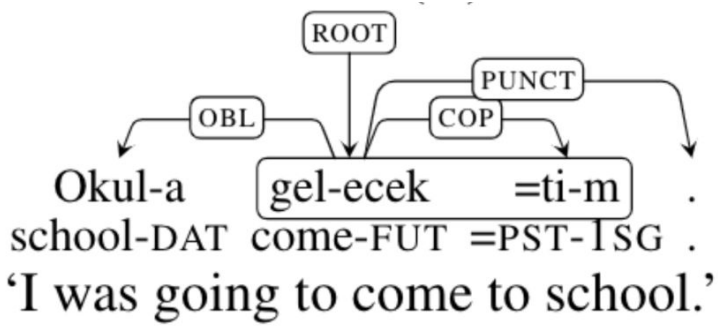
then, what about clitic copulars?

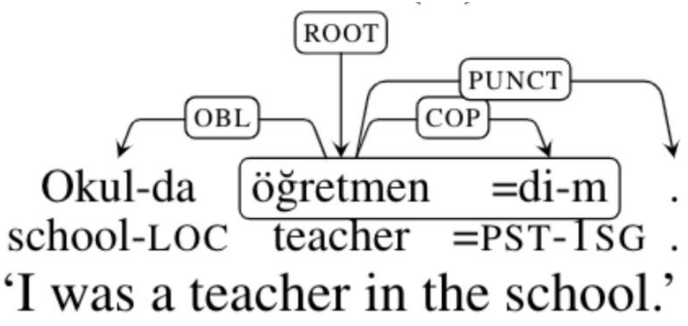




issue #2: copula

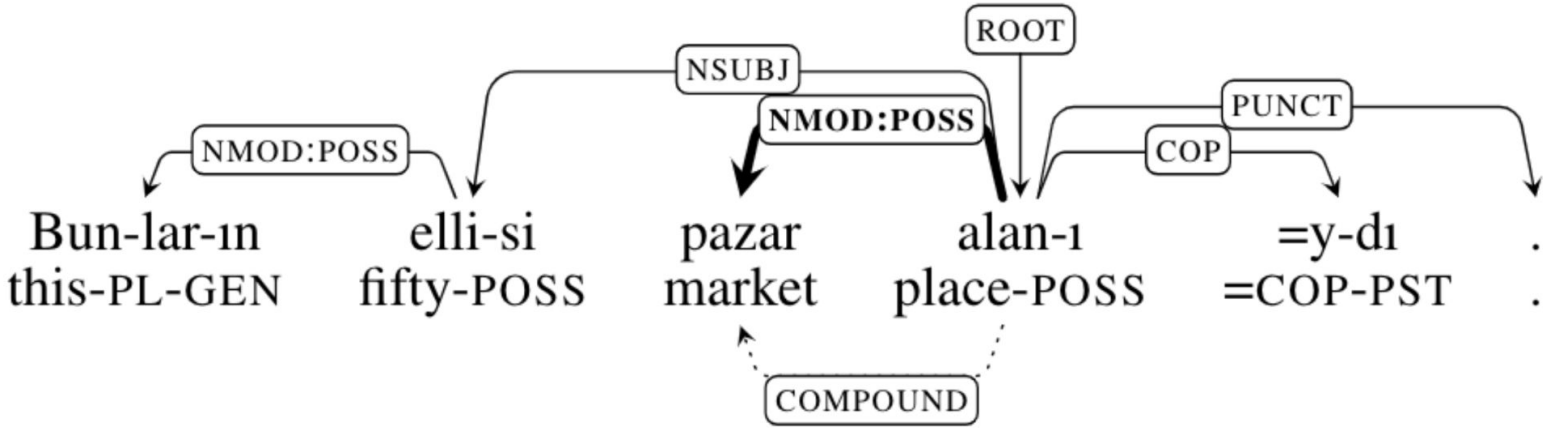
zero copula?





issue #3: compound

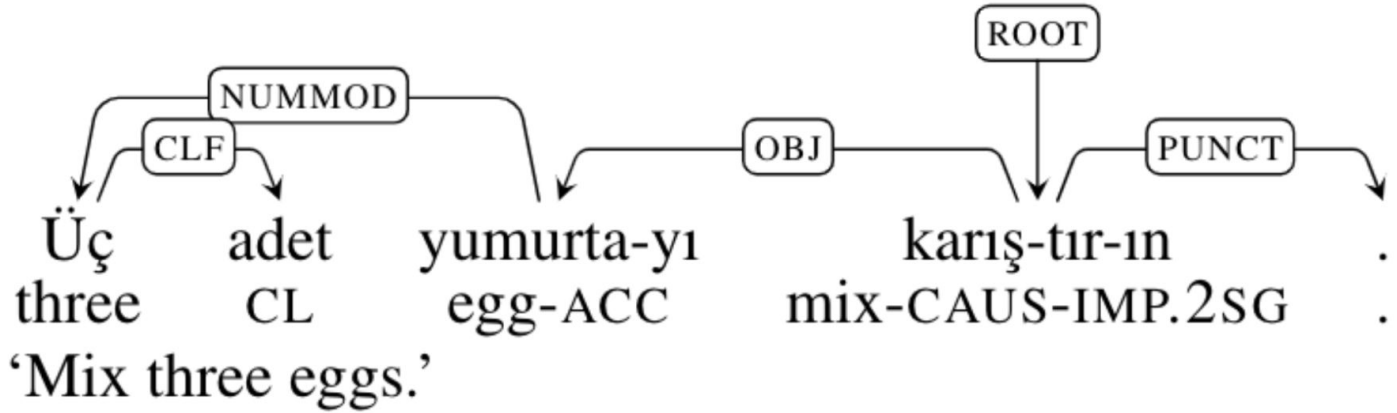
definition is clear, application is all over the place



‘50 of these were marketplaces.’

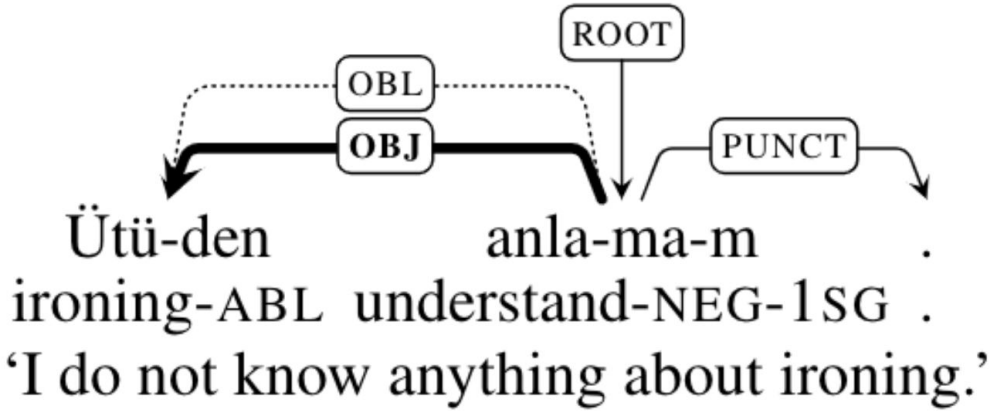
issue #4: classifier

previous annotations of “adet” or “tane” is not unified



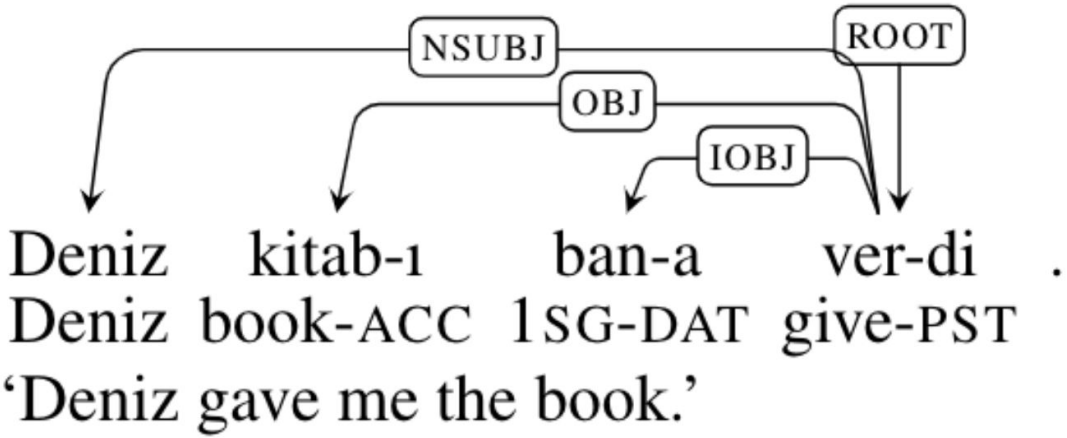
issue #5: core arguments

arguments with lexical cases used to be grouped with adjuncts



issue #5: core arguments

indirect objects as well.



experiments

- **task: predicting syntactic attachment (UAS) and successfully determining syntactic and morphological tags (LAS)**
- Stanford's neural parser (Dozat et al., 2017; Kanerva et al., 2018)
- unidirectional LSTM for word embeddings
- bidirectional LSTM for head-dependency relations
- randomly assigned training (80%), development (10%), and test (10%)

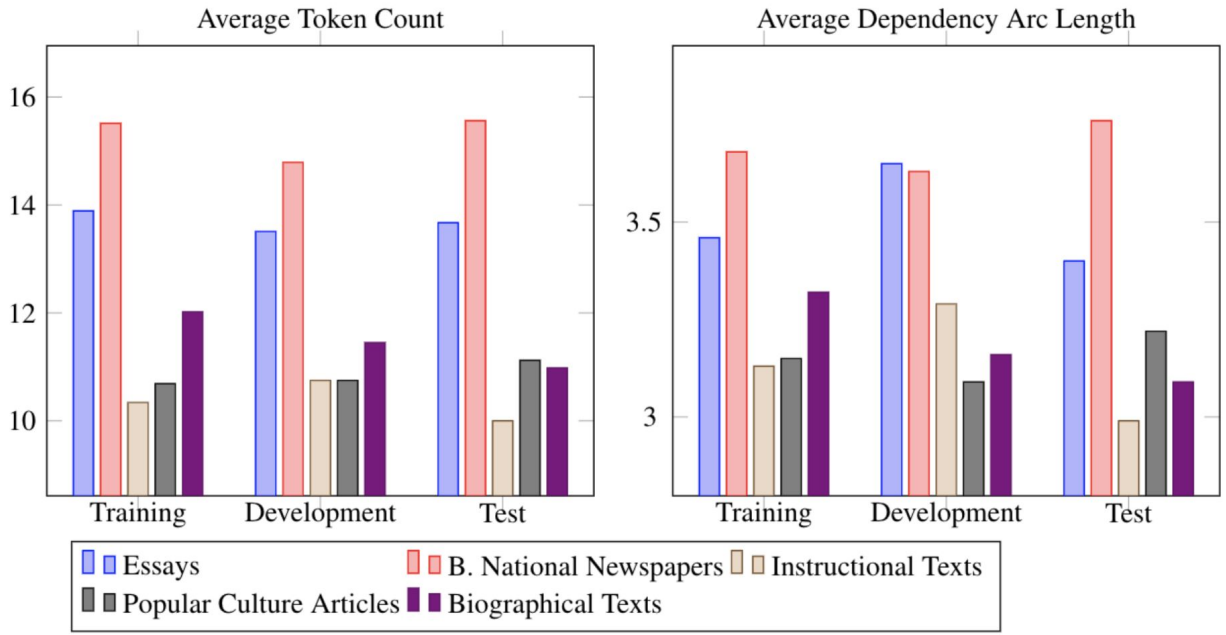
experiments: BOUN

predictive parsing results are generally uniform, except for essays. but why?

Genre	UAS F1-Score	LAS F1-Score
Essays	68.73	59.18
National Newspaper	81.59	76.04
Instructional Text	79.22	72.65
Popular Culture	77.69	71.13
Biography	80.28	73.68
Total	**77.36**	**70.37**

experiments: BOUN

*maybe, **number of words** or **locality**? not supported.*



experiments: pooled

training size and training data quality seems effective

Training set	Training size	Test set	Test size	UAS F1-score	LAS F1-score
IMST-UD	3,685	BOUN	979	69.38	58.65
BOUN	7,803	BOUN	979	77.36	70.37
BOUN+IMST-UD	11,488	BOUN	979	77.57	70.50
IMST-UD	3,685	IMST-UD	975	75.49	65.53
BOUN	7,803	IMST-UD	975	73.63	62.92
BOUN+IMST-UD	11,488	IMST-UD	975	76.86	66.79
IMST-UD	3,685	PUD	1,000	65.28	49.50
BOUN	7,803	PUD	1,000	72.33	59.57
BOUN+IMST-UD	11,488	PUD	1,000	72.76	60.39
IMST-UD	3,685	BOUN+IMST-UD	1,954	71.89	61.62
BOUN	7,803	BOUN+IMST-UD	1,954	75.67	66.99
BOUN+IMST-UD	11,488	BOUN+IMST-UD	1,954	77.25	68.82
IMST-UD	3,685	BOUN+IMST-UD+PUD	2,954	69.03	56.37
BOUN	7,803	BOUN+IMST-UD+PUD	2,954	74.22	63.78
BOUN+IMST-UD	11,488	BOUN+IMST-UD+PUD	2,954	75.30	65.17

takes

- discrepancy in parsing results **are not due to** average token count and arc length
  - maybe, due to the nature of the texts?
- more data provide better parsing results
- linguistically-adequate annotations enhance treebank quality, thus parsing results

references

Aksan Y., Aksan M., Koltuksuz A., Sezer T., Mersinli Ü., Demirhan U. U., Yılmazer H.,Atasoy G., Öz S., Yıldız İ., Kurtoğlu Ö. (2012) Construction of the Turkish NationalCorpus (TNC). In:Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA),Istanbul, Turkey, pp. 3223–3227, URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/991_Paper.pdf

De Marneffe, M.-C., & Nivre, J. (2019). Dependency Grammar. Annual Review of Linguistics, 5(1), 197–218. doi:10.1146/annurev-linguistics-011718-011842

Dozat T., Qi P., Manning C. D. (2017) Stanford's graph-based neural dependency parser atthe CoNLL 2017 shared task. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pp. 20–30

Kanerva J., Ginter F., Miekka N., Leino A., Salakoski T. (2018) Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In:Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics, Brussels, Belgium, pp. 133–142, URL http://www.aclweb.org/anthology/K18-2013

Mel'cuk I. 1988. Dependency Syntax: Theory and Practice. Albany, NY: SUNY Press

Milicevic J. 2006. A short guide to the Meaning–Text linguistic theory. J. Koralex 8:187–233

Nivre J., de Marneffe M.-C., Ginter F., Goldberg Y., Hajic J., Manning C. D., McDon-ald R., Petrov S., Pyysalo S., Silveira N., Tsarfaty R., Zeman D. (2016) Universal De-pendencies v1: A multilingual treebank collection. In:Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, pp. 1659–1666, URL https://www.aclweb.org/anthology/L16-1262

Sak H., Güngör T., Saraçlar M. (2011) Resources for Turkish morphological processing.Language Resources and Evaluation 45 (2): 249–261

Tesniere L. 1959. Eléments de syntaxe structurale. Paris: Ed. Klincksieck