Improving the Annotations in the Turkish Universal Dependency Treebank

Utku Türk[‡], Furkan Atmaca[‡], Şaziye Betül Özateş[*], Balkız Öztürk[‡], Tunga Güngör[*], Arzucan Özgür[*]

[‡]Department of Linguistics
[*]Department of Computer Engineering
Boğaziçi University

August 30, 2019

Improving Turkish UD

Path

- Previous Studies on Turkish UD
- Process
- Revisions
- Experiments
- Conclusion and Future Work

Aim

- To unify annotation patterns and decisions within Turkish UD
- To create a basis for future treebanks

Previous Studies on Turkish UD

	IMST-UD	**PUD**	**GB**
Word Count	56.396	16.535	*unknown*
Copyright	CC-BY-NC-SA	CC-BY-SA 3.0	CC BY-SA 4.0
UD version	v2.2	v2.2	v2.4
Content	Non-Fiction, Newspaper	Newspaper, Wikipedia	Examples from a Grammar book
Method	manual annotation in non-UD, semi-automatic conversion	manual annotation in non-UD, automatic conversion	manual annotation in non-UD, automatic conversion
Syntactic Relation Number	32	41	*unkown*

Table 1: State of Turkish UD Treebanks

IMST-UD Under the Microscope

Timeline

- **MST**: Atalay et al. (2003) and Oflazer et al. (2003)
- **IMST**: Sulubacak, Pamay, et al. (2016)
- **IMST-UD**: Sulubacak, Gökırmak, et al. (2016)

IMST-UD Under the Microscope

Problems

- Lack of a linguistic team behind the annotation process
- Failure to check annotations
- As an immediate result, an incoherent picture after the automatic conversion
- Due to morpho-ortographic similarities, the automatic conversion process failed to capture the intriguing nature of Turkish.
  - **Embedded Clause Structure** and **Possessive marked nominal compounds**
  - **Embedded Subjects** and **Genitive Marking**
  - Core Arguments with **lexically-driven** case marking

IMST-UD Under the Microscope

Problems

- Lack of a linguistic team behind the annotation process
- Failure to check annotations
- As an immediate result, an incoherent picture after the automatic conversion
- Due to morpho-ortographic similarities, the automatic conversion process failed to capture the intriguing nature of Turkish.
  - **Embedded Clause Structure** and **Possessive marked nominal compounds**
  - **Embedded Subjects** and **Genitive Marking**
  - Core Arguments with **lexically-driven** case marking

IMST-UD Under the Microscope

Problems

- Lack of a linguistic team behind the annotation process
- Failure to check annotations
- As an immediate result, an incoherent picture after the automatic conversion
- Due to morpho-ortographic similarities, the automatic conversion process failed to capture the intriguing nature of Turkish.
  - **Embedded Clause Structure** and **Possessive marked nominal compounds**
  - **Embedded Subjects** and **Genitive Marking**
  - Core Arguments with **lexically-driven** case marking

IMST-UD Under the Microscope

Problems

- Lack of a linguistic team behind the annotation process
- Failure to check annotations
- As an immediate result, an incoherent picture after the automatic conversion
- Due to morpho-ortographic similarities, the automatic conversion process failed to capture the intriguing nature of Turkish.
  - **Embedded Clause Structure** and **Possessive marked nominal compounds**
  - **Embedded Subjects** and **Genitive Marking**
  - Core Arguments with **lexically-driven** case marking

IMST-UD Under the Microscope

Problems

- Lack of a linguistic team behind the annotation process
- Failure to check annotations
- As an immediate result, an incoherent picture after the automatic conversion
- Due to morpho-ortographic similarities, the automatic conversion process failed to capture the intriguing nature of Turkish.
  - **Embedded Clause Structure** and **Possessive marked nominal compounds**
  - **Embedded Subjects** and **Genitive Marking**
  - Core Arguments with **lexically-driven** case marking

IMST-UD Under the Microscope

Problems

- Lack of a linguistic team behind the annotation process
- Failure to check annotations
- As an immediate result, an incoherent picture after the automatic conversion
- Due to morpho-ortographic similarities, the automatic conversion process failed to capture the intriguing nature of Turkish.
  - **Embedded Clause Structure** and **Possessive marked nominal compounds**
  - **Embedded Subjects** and **Genitive Marking**
  - Core Arguments with **lexically-driven** case marking

IMST-UD Under the Microscope

Problems

- Lack of a linguistic team behind the annotation process
- Failure to check annotations
- As an immediate result, an incoherent picture after the automatic conversion
- Due to morpho-ortographic similarities, the automatic conversion process failed to capture the intriguing nature of Turkish.
  - **Embedded Clause Structure** and **Possessive marked nominal compounds**
  - **Embedded Subjects** and **Genitive Marking**
  - Core Arguments with **lexically-driven** case marking

Process

Team

3 Linguist and 3 NLP specialist

Workflow

- Settling on the definitions of UD syntax and providing examples for every possible syntactic relation
- Recording inconsistencies and erroneous tagging in the IMST-UD
- Cross checking reports
- Discussions regarding problematic cases

Process

Team

3 Linguist and 3 NLP specialist

Workflow

- Settling on the definitions of UD syntax and providing examples for every possible syntactic relation
- Recording inconsistencies and erroneous tagging in the IMST-UD
- Cross checking reports
- Discussions regarding problematic cases

Revisions

- Transparency of **Embedded** Clauses
- Syntactic **Category** Errors
- Representation of **Core** Arguments

Type	IMST-UD	BIMST-UD	$n_{alternations}$
Embedded	NMOD	ADVCL	666
	OBJ	CCOMP	481
	NMOD:POSS	NSUBJ	312
Category	OBJ	NSUBJ	276
Core	OBL	IOBJ	194
	OBL	OBL:ARG	137

Table 2: The number of alternations for the most frequent changes

Revisions

Embedded Structures

Turkish makes use of nominalization processes in all its embedded structures. While embedded structures that are core arguments of a verb show nominal properties, relative clauses are complex adjectival constructions.

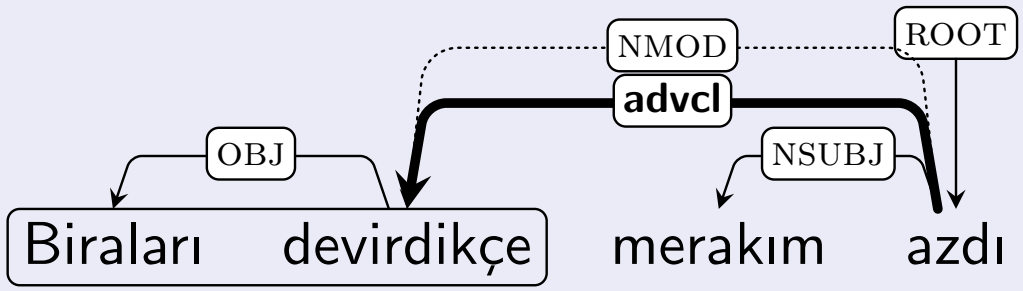
(1) Sen-**in** de bu gösteri-yi izle-me-**ni** iste-r-di-m.
you-GEN too this show-ACC watch-NMLZ-POSS want-AOR-PST-1SG
"I would have wanted you to watch this show, as well."

(2) Sen-**in** elma-**nı** iste-r-di-m.
you-GEN apple-POSS want-AOR-PST-1SG
"I would have wanted your apple."

Revisions

Embedded Structures: NMOD > ADVCL

(3)



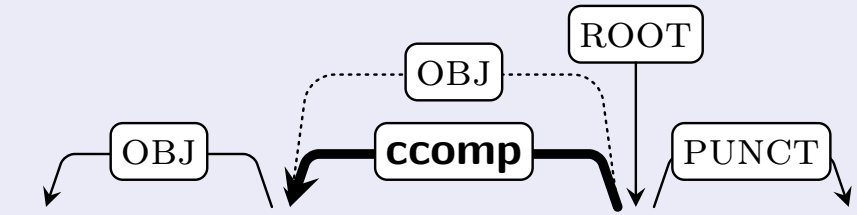
Bira-lar-ı devir-dik-çe merak-ım az-dı.
beer-PL-ACC topple-NMLZ-CVB curiosity-POSS.1SG get.wild-PST

"As I finished my beers, my curiosity peaked."

Revisions

Embedded Structures: OBJ > CCOMP

(4)



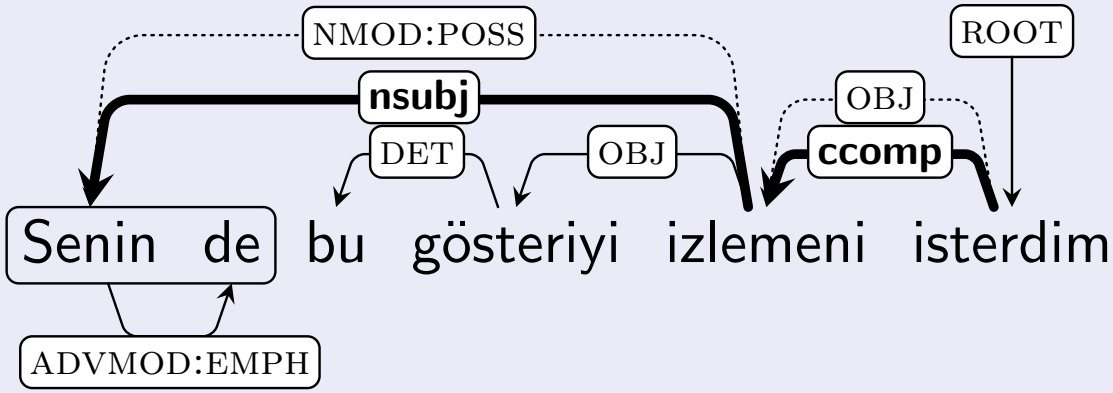
Ne di-yeceğ-im-i bil-e-me-di-m.
what say—FUT-POSS-ACC know--NEG-PST-1SG.

"I didn't know what to say."

Revisions

Embedded Structures: NMOD:POSS > NSUBJ

(5)



Sen-in de bu gösteri-yi izle-me-ni iste-r-di-m.
you-GEN too this show-ACC watch-NMLZ-POSS want-AOR-PST-1SG

"I would have wanted you to watch this show, as well."

Revisions

Syntactic Category Errors

- We can group these errors under two headings
  - Effects of bare objects

(6) Çorap ör-dü-m
sock.ACC knit-PST-1SG
"I did sock-knitting."

  - Effects of possessive marker

(7) Göz-ler-i parla-dı.
eye-PL-POSS sparkle-PST
"Her eyes sparkled."

(8) Oje-si parla-dı.
nailpolish-POSS sparkle-PST
"Her nailpolish sparkled."

Revisions

Syntactic Category Errors

- We can group these errors under two headings
  - Effects of bare objects

    (9) Çorap ör-dü-m
    sock.ACC knit-PST-1SG
    "I did sock-knitting."

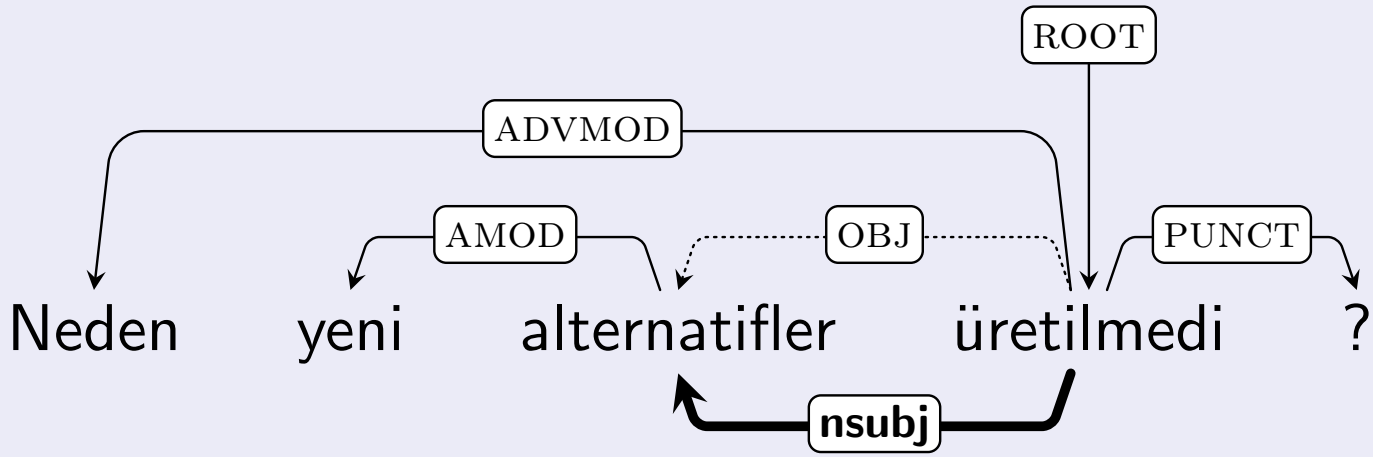
  - Effects of possessive marker

    (10) Göz-ler-i parla-dı.
    eye-PL-POSS sparkle-PST
    "Her eyes sparkled."

    (11) Oje-si parla-dı.
    nailpolish-POSS sparkle-PST
    "Her nailpolish sparkled."

Syntactic Category Errors

- We can group these errors under two headings
  - Effects of bare objects



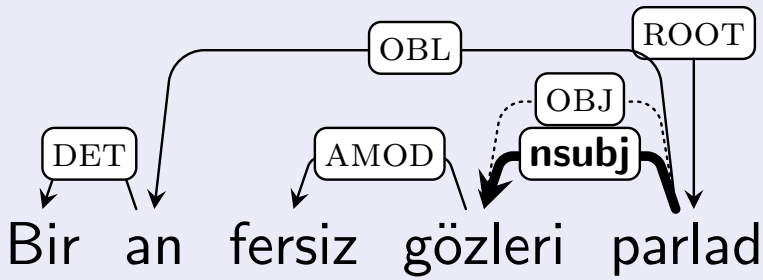
(12)

Neden yeni alternatif-ler üret-il-me-di?
why new alternative-PL produce-PASS-NEG-PST?

"Why weren't new alternatives produced?"

Syntactic Category Errors

- We can group these errors under two headings
  - Effects of possessive marker



(13)

Bir an fersiz göz-ler-i parla-dı.
one moment dull eye-PL-POSS sparkle-PST

"Her dull eyes sparkled for a moment."

Revisions

Core Arguments

In addition to the accusative case, Turkish can also mark its core arguments with cases that are generally used to mark adjuncts, such as locative, ablative, comitative, dative.

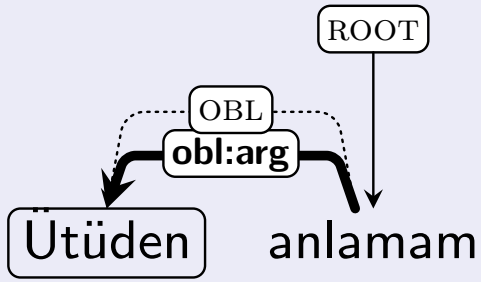
(14) Ütü-**den** anla-ma-m
iron-ABL know-NEG-1SG
"I do not know a thing about ironing."

(15) Kitap-**tan** sayfa-lar dökül-üyor.
book-ABL page-PL fall-PROG.
Pages are falling out of this book.

Revisions

Core Arguments: OBL > OBL:ARG

(16)



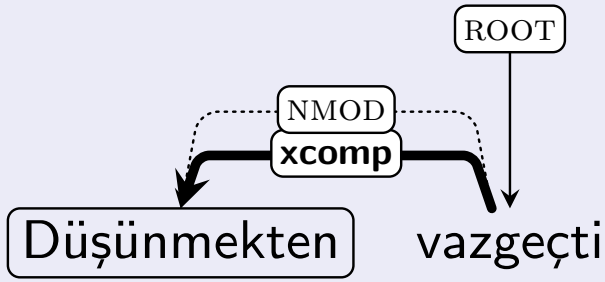
Ütü-den anla-ma-m.
ironing-ABL understand-NEG-1SG

"I do not know a thing about ironing."

Revisions

Core Arguments: NMOD > XCOMP

(17)



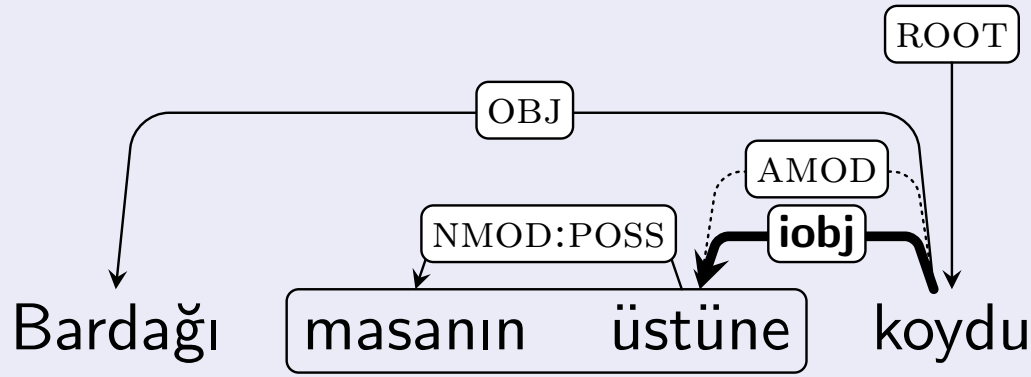
Düşün-mek-ten vazgeç-ti.
thinking-NMLZ-ABL give.up-PST

"She decided not to think about it."

Revisions

Introduced Dependencies: IOBJ

(18)



Bardağ-ı masa-nın üst-ün-e koy-du.
glass-ACC table-GEN top-POSS-DAT put-PST

"She put the cup on the table."

Experiment

Parser Information

- Training on both a transition-based LSTM dependency parser (Özateş et al., 2018) and a graph-based neural parser (Dozat et al., 2017).
- Both projective and non-projective dependencies were included.
- Turkish word embeddings of the CoNLL-17 pre-trained word embeddings from Ginter et al. (2017) were used.
- We used labeled and unlabeled attachment score metrics.

Experiment

Results

		IMST-UD	BIMST-UD
Transition-based parser	**UAS**	65.91	**68.66**
	LAS	59.06	58.98
Graph-based parser	**UAS**	71.55	**75.49**
	LAS	64.86	**65.53**

Table 3: UAS and LAS scores of the two parsers on the previous and updated versions of the IMST-UD treebank.

Conclusion

- Finding correct head-dependent relations ↑
- Labeling efficiency in graph-based parser ↑

∴ We found the following: (i) a linguistically fine-grained analysis improved parser performance, (ii) making adjustments according to the needs of a specific language does not always compromise parsing performance.

Thank you!

Acknowledgement

This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under grant number 117E971 and as a graduate scholarship.

Selected References

Atalay, N. B., K. Oflazer, and B. Say (2003). "The annotation process in the Turkish treebank". In: *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003* (cit. on p. 4).

Çöltekin, Ç. (2015). "A Grammar-Book Treebank of Turkish". In: *Proceedings of the 14th Workshop on Treebanks and Linguistic Theories (TLT 14)*. Ed. by M. Dickinson, E. Hinrichs, A. Patejuk, and A. Przepiórkowski. Warsaw, Poland, pp. 35–49.

Dozat, T., P. Qi, and C. D. Manning (2017). "Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task". In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 20–30 (cit. on p. 27).

Ginter, F., J. Hajič, J. Luotolahti, M. Straka, and D. Zeman (2017). *CoNLL 2017 Shared Task - Automatically Annotated Raw Texts and Word Embeddings*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University (cit. on p. 27).

Göksel, A. and C. Kerslake (2005). *Turkish: A Comprehensive Grammar*. Comprehensive grammars. Routledge.

Nivre, J., M.-C. De Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, et al. (2016). "Universal Dependencies v1: A Multilingual Treebank Collection". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 1659–1666.

Oflazer, K., B. Say, D. Z. Hakkani-Tür, and G. Tür (2003). "Building a Turkish Treebank". In: *Treebanks, Building and Using Parsed Corpora*, pp. 261–277 (cit. on p. 4).

Özateş, Ş. B., A. Özgür, T. Güngör, and B. Öztürk (2018). "A Morphology-based Representation Model for LSTM-based Dependency Parsing of Agglutinative Languages". In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 238–247 (cit. on p. 27).

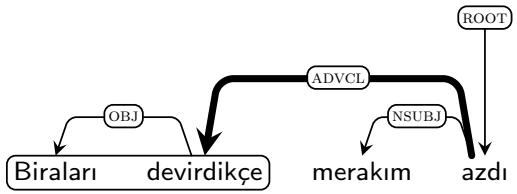
Sulubacak, U., M. Gökırmak, and F. M. Tyers (2016). "Universal Dependencies for Turkish". In: *Proceedings of the 26th International Conference on Computational Linguistics (COLING-16)*, pp. 3444–3454 (cit. on p. 4).

Sulubacak, U., T. Pamay, and G. Eryiğit (2016). "IMST: A Revisited Turkish Dependency Treebank". In: *In Proceedings of 1st International Conference on Turkic Computational Linguistics, TurCLing*, pp. 1–6 (cit. on p. 4).

Tyers, F. M., J. Washington, Ç. Çöltekin, and A. Makazhanov (2017). "An Assessment of Universal Dependency Annotation Guidelines for Turkic Languages". In: Tatarstan Academy of Sciences.

Appendix: Intorduced Dependencies I

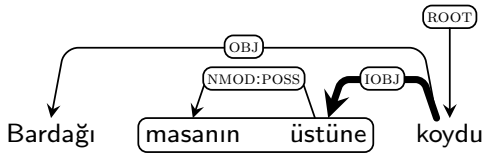
(19) `advcl`:



Bira-lar-ı devir-dik-çe merak-ım az-dı.
beer-PL-ACC topple-NMLZ-CVB curiosity-POSS.1SG get.wild-PST

"As I finished my beers, my curiosity peaked."

(20) `iobj`:

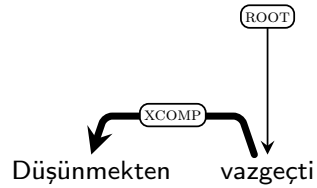


Bardağ-ı masa-nın üst-ün-e koy-du.
glass-ACC table-GEN top-POSS-DAT put-PST

"She put the cup on the table."

Appendix: Intorduced Dependencies II

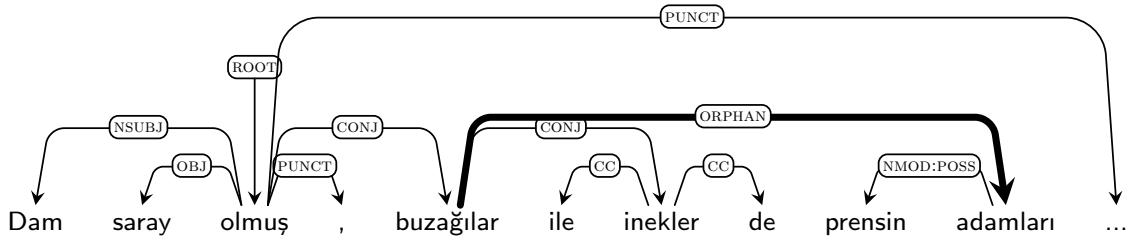
(21) `xcomp`:



Düşün-mek-ten vazgeç-ti.
thinking-NMLZ-ABL give.up-PST

"She gave up thinking."

(22) `orphan:`

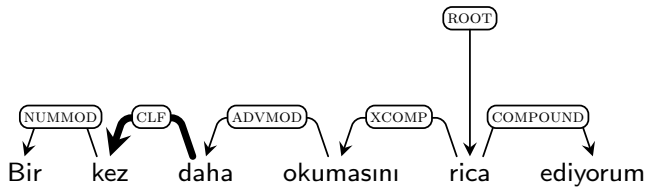


Dam saray ol-muş, buzağı-lar ile inek-ler de prens-in adam-lar-ı...
stable palace be-EVD calf-PL with cow-PL too prince-GEN man-PL-POSS...

"Apparently stables became palaces, and cows with their calves henchman of the prince..."

Appendix: Intorduced Dependencies III

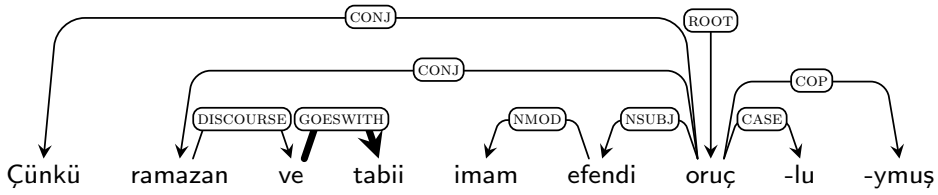
(23) `clf`:



Bir kez daha oku-ma-sın-ı rica ed-iyor-um.
one time more read-NMLZ-POSS-ACC request do-PROG-1SG

"I request that he/she reads it one more time."

(24) `goeswith`:

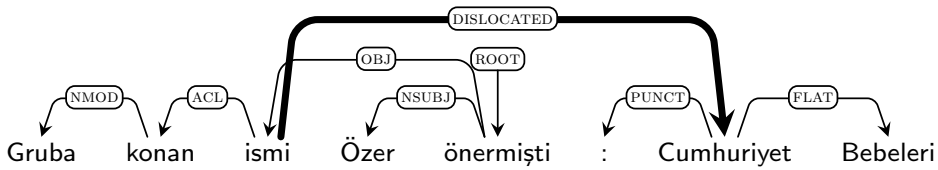


Çünkü ramazan ve tabii imam efendi oruç-lu-ymuş.
because Ramadan and ofcourse imam mister fast-COM-COP

"Because it is Ramadan and, of course, our dear imam is fasting."

Appendix: Intorduced Dependencies IV

(25) `dislocated`:



Grub-a	kon-an	ism-i	Özer	öner-miş-ti:	Cumhuriyet	Bebe-ler-i.
band-DAT	put-NMLZ	name-POSS	Özer	suggest-EVD-PST:	Republic	baby-PL-POSS

"It was Özer who suggested the name for the band: The Children of the Republic."