First Steps towards Universal Dependencies for Laz

Utku Türk[‡], Kaan Bayar[‡], Ayşegül Dilara Özercan[‡], Görkem Yiğit Öztürk[‡], Şaziye Betül Özateş[*]

[‡]Department of Linguistics, [*]Department of Computer Engineering, Boğaziçi University, Bebek, 34342 İstanbul, Turkey
{utku.turk, kaan.bayar, aysegul.ozercan, gorkem.ozturk, saziye.bilgin}@boun.edu.tr

Laz Background

- Endangered South Caucasian language (< 500,000 speakers).
- Mainly spoken in North East Turkey.
- Exhibits extensive dialectical variation.
- Agglutinative language with context-sensitive SOV word order.

The BOUN Laz Treebank Details

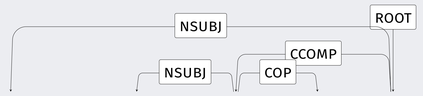
- Based on **Atina-Pazar** dialect of Laz, yet includes sentences from **Ardeshen** dialect as well.
- 576 sentences, 2306 tokens. We aim to have 1500 sentences.
- Consists of two parts: Articles Grammar Book
- The full list of resources is as follows: Emgin [7], Demirok [3], Demirok et al. [5], Demirok [2], Demirok [4], Demirok [1], Öztürk [8], Öztürk and Taylan [11], Öztürk [9], Öztürk and Pöchtrager [10].
- Available: `https://github.com/boun-tabi/UD_Laz-BOUN`
- A team of four linguists and an NLP specialist.

Linguistic Decisions #1: Person Marking

(1) OBL.1SG/2SG > P.1SG/2SG > A.1SG > OBL.3SG = P.3SG = A.2SG/3SG

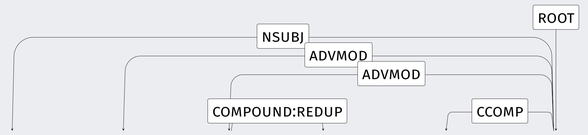
- Explicit marking of the person depends on the hierarchy in (1).
- The person and the number marked in the verb does not always points to the controller with the same theta-role.
- Existing UD-solution: Basque uses case layers for these features as the following: `Person[erg]={1,2,3}`.
- However, case-controller relation is also not one-to-one in Laz.
- **Our proposal**: Usage of case layer for canonical examples and usage of dependency relation (`Person[nsubj]={1,2,3}`) as a layer for non-canonical examples.

Linguistic Decisions #2: YA-sentences



(2) Tanura-k Şana noseri on ya
Tanura-ERG Şana.NOM smart be.PRS COMP
'Tanurak thinks/says that Şana is smart.'

- YA may be existentially closed; no need for a verb like *say or think*.
- The meaning of YA sentences are context sensitive.
- A possible analysis: YA-sentences are 'opinion' sentences similar to *as for Tanura* or *Tanura'ya göre*.
- Problematic sentences as in (3).
- 'Arte thought that he would swim silently yesterday' should be possible. But it is not acceptable to native speakers.



(3) Arte-k ğoma uneneli uneneli vinçirare ya
Arte-ERG yesterday silent silent swim.1SG.FUT COMP
'Yesterday, Arte silently said that he would swim.'

- **Our analysis:** YA is a complex lexical item that encapsulates complementizer and a verb.
- This enables a possible modification of the event, unlike the previous analysis.

[1] Ö. Demirok. Agree as a unidirectional operation: Evidence from Pazar Laz. Master's thesis, Boğaziçi University, 2013.
[2] Ö. Demirok. The status of roots in event composition: Laz. *Lingue e linguaggio, Rivista semestrale*, (1/2014):83–102, 2014. ISSN 1720-9331. doi: 10.1418/77001.
[3] Ö. Demirok. A modal approach to dative subjects in Laz. In S. Hucklebridge and M. Nelson, editors, *NELS 48: Proceedings of the Forty-Eighth Annual Meeting of the North East Linguistic Society*. CreateSpace Independent Publishing Platform, 2018.
[4] Ö. Demirok. Non-linear blocking of portmanteaus: a case study on Laz. Talk given at NanoLAB, Masaryk University, Brno, 2020.
[5] Ö. Demirok, D. Özyıldız, and B. Öztürk. Complementizers with attitude. In M. Baird, editor, *NELS 49: Proceedings of the Forty-Ninth Annual Meeting of the North East Linguistic Society: Volume 3*. Amherst, MA: GLSA, Dept. of Linguistics, 2019.
[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.
[7] B. Emgin. *Finiteness and complementation in Laz*. PhD thesis, Boğaziçi University, 2009.
[8] B. Öztürk. Applicatives in Pazar Laz. Talk given at The South Caucasian Chalk Circle 3, Paris, 2016.
[9] B. Öztürk. The loss of case system in Ardeshen Laz and its morphosyntactic consequences. *STUF - Language Typology and Universals*, 72(2):193 – 219, 2019. doi: https://doi.org/10.1515/stuf-2019-0008.
[10] B. Öztürk and M. A. Pöchtrager. *Pazar Laz*. Lincom Europa München, 2011.
[11] B. Öztürk and E. E. Taylan. Omnipresent little v in Pazar Laz. Talk given at Little v Workshop, University of Leiden, 2013.

Linguistic Decisions #3: Affirmative Preverbs

(4) Ali ko-mo-xt'-u.
Ali.NOM PV_{aff}-PV_{spat}-come-PST.3SG
'Ali certainly came.' [10]

- PV_{aff} in (4) is ambiguous between habitual or certainty reading.
- Evidence-related `Evident` feature is the closest UD candidate.
- However, (4) is grammatical even when the speaker does not witness the event. How about `Polarity=Aff`?
- Affirmative reading is still possible without the preverb.
- Other values for `Polarity` are not directly related to the phenomenon in Laz.
- **Our solution:** A new value as `Aspect=Crt`. When distinguishable: `Aspect=Hab`.

Experiments

- We used a multilingual parser `UDify` [6] with no training set.
- The reasons behind: Having a small data set and cross-linguistic comparability with other un(der)represented languages.

Treebank	N_{Token}	UAS	LAS
BOUN Laz Treebank	2K	44.15	29.05
Akkadian-PISANDUB	1K	27.65	4.54
Amharic-ATT	5K	17.38	3.49
Cantonese-HK	6K	46.82	32.01
Erzya-JR	15K	31.90	16.38
Komi Zyrian-IKDP	1K	36.01	22.12
Komi Zyrian-Lattice	2K	28.85	12.99
Naija-NSC	12K	45.75	32.16
Sanskrit-UFAL	1K	40.21	18.56
Warlpiri-UFAL	< 1K	21.66	7.96
Yoruba-YTB	2K	37.62	19.09

- Unlike some of the languages, Laz do not have any related language known to `UDify`.
- We did not fine-tune the `UDify`.
- Got close results to Naija (English creole) and Cantonese (belongs to Sino-Tibetan language family) with less data and no related languages.
- Low token per sentence has an effect on our relatively high score.