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

Abstract

This paper presents the first treebank for the Laz language, which is also the first Universal Dependencies Treebank for a South Caucasian language. This treebank aims to create a syntactically and morphologically annotated resource for further research. We also aim to document an endangered language in a systematic fashion within an inherently cross-linguistic framework: the Universal Dependencies Project (UD). As of now, our treebank consists of 576 sentences and 2,306 tokens annotated in light with the UD guidelines. We evaluated the treebank on the dependency parsing task using a pretrained multilingual parsing model, and the results are comparable with other low-resourced treebanks with no training set. We aim to expand our treebank in the near future to include 1,500 sentences. The bigger goal for our project is to create a set of treebanks for minority languages in Anatolia.

1 Introduction

In recent years, many understudied languages have been in the spotlight of NLP studies. Within the Universal Dependencies Framework (Nivre et al., 2016), languages like Wolof (Dione, 2019), Mbyá Guaraní (Thomas, 2019), Eryza (Rueter and Tyers, 2018), Bhojpuri (Ojha and Zeman, 2020), and many others have been introduced to NLP studies. Laz is also another understudied language on which there are no NLP resources, with the exception of a recent morphological analyzer (Önal and Tyers, 2019).

Laz is spoken in the Southeastern part of the Black Sea among a declining population which is estimated to consist of somewhere between 250,000 and 500,000 people (Haznedar, 2018). It is a highly agglutinative language which makes use of prefixes and suffixes, and it is reported to have 16 different slots for verb inflection (Öztürk and Pöchtrager, 2011). Laz is also reported to have extensive dialectal variation (Öztürk and Pöchtrager, 2011). In this work, when we use the word Laz, we specifically mean the Atina-Pazar dialect of Laz.

Through this work, we hope to contribute to the revitalization efforts of the Laz language by providing a gold standard dependency treebank. We provide the first publicly available human-annotated morphosyntactic Laz treebank. We utilize the already existing Universal Dependency framework to represent UPOS tags, morphological features, and syntactic dependency relations. We also provide parsing results with no training data using UDify (Kondratyuk and Straka, 2019). Thus, we provide attachment scores (UAS and LAS) which are comparable with other low resource languages within the UD framework.

2 BOUN Laz Treebank

2.1 Data and Treebank Statistics

The data used in this treebank consist of linguistic examples from academic works which describe the Laz language. These works include theses, articles, proceedings, and presentation handouts.[1] In our

[1]The full list of resources is as follows: Emgin (2009), Demirok (2018), Demirok et al. (2019), Demirok (2014), Demirok (2020), Demirok (2013), Öztürk (2016), Öztürk and Taylan (2013), Öztürk (2019)

Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020), pages 189–194
Barcelona, Spain (Online), December 13, 2020

future work, we want to include reference grammars, user-generated data, and folk stories. As of now, we present 576 sentences and 2,306 tokens which have glosses in the source files they are taken, but not syntactically annotated before. We annotated every sentence and every token manually following UD guidelines. This process includes annotating morphological features, syntactic dependency relations and POS tags. In our treebank, we also included a gloss and English translation for every sentence. Additionally, we used easily parsable sentence IDs which follow the `source-genre_number` template. When we include sentences extracted from fictions, we will also provide paragraph ID for sentences.

Caucasian languages have always presented a challenge to mainstream notion in linguistics research. The challenge mainly stem from their highly allomorphic and fusional morphology (Blix, 2020; Demirok, 2020). We believe that the addition of these languages would pose new questions for the typological adequacy of the UD framework. These issues include the representation of complex subordinating conjunctions and the need for finer analysis in the morphological representation.

2.2 Annotation Process

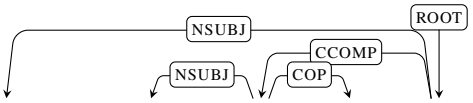
The treebank was annotated with a team of 4 linguists, comprised of 3 annotators and one reviewer. The annotators were responsible for their own batch of sentences. After the annotation, these sentences were sent to the reviewer to be checked. All changes made in this review process were discussed by the entire annotator group. The decision agreed upon was then applied uniformly to the treebank and recorded as part of the guidelines we prepared.

3 Linguistic Analysis of Laz Dependency Treebank

As we stated earlier, we followed UD v2 guidelines to annotate our Laz treebank. There were some challenging data that needed to be implemented to UD framework. We mention three of them in this section with the linguistic discussion behind our decision.

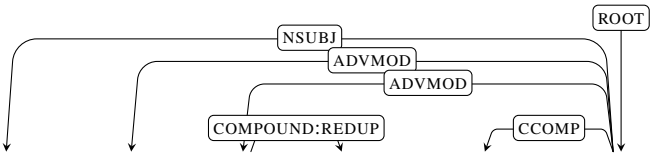
3.1 Complementizer 'YA'

One challenge in annotating Laz linguistic data is coding sentences that include the complementizer 'ya'. Interestingly, embedded YA sentences may be existentially closed, meaning that they do not need any verb to be reported as in (1) (Demirok et al., 2019).[2]



(1) Tanura-k Şana noseri on ya
Tanura-ERG Şana.NOM smart be.PRS COMP
'Tanurak thinks/says that Şana is smart.' (Adapted from Demirok et al. (2019))

The problem with such sentences is that they lack an attitude verb even though they contain a context-sensitive meaning of reporting. Since UD guidelines only allow empty nodes for elision and conjunction, we were not able to implement a solution which involves speculating a hidden verb (Droganova and Zeman, 2019). A syntax and semantics oriented explanation of this phenomenon pushes for a complex YA analysis in which the YA complementizer encompasses the attitude verb, *say* in this case. This analysis is supported with other examples such as (2). The adverbials *slowly* and *yesterday* modifies the omitted main verb *say* instead of the embedded verb *swim*.



(2) Arte-k ğoma uneneli uneneli vinçirare ya.
Arte-ERG yesterday silent silent swim.1SG.FUT COMP

[2] 1 = first person, 2 = second person, 3 = third person, A = agent, AUG = augmentative, AUX = auxiliary, CAUS = causative, COMP = complementizer, COND = conditional, ERG = ergative, FUT = future, IMPF = imperfective, INTR = intransitive, NOM = nominative, OBL = oblique, P = patient, PL = plural, PRF = perfect, PRS = present, PST = past, PV = preverb, SBJV = subjunctive, SG = singular, TR = transitive, TS = thematic suffix.

'Yesterday, Arte silently said that he would swim.'

Following the discussion in the UD repository,[3] we marked the complementizer YA as the root of the sentence. One other possibility is to treat the ergative case as the marker for '*in somebody's opinion*'. With '*in my opinion*' reading, the sentence (1) is grammatical and means "According to Tanura, Şana is smart." However, sentence (2) would not be grammatical and we would still need YA-as-a-root solution. Thus, we did not use '*in somebody's opinion*' reading in our annotations even though this is a possible reading.

3.2 Morphological Person Marking

Another challenging aspect is the morphological annotation of person and number agreement. The verbal agreement paradigm in Laz makes use of both suffixes and prefixes. Verbs may host agreement markers not only for subjects, but also objects, indirect objects, and non-core arguments such as benefactors. However, these markers are partially co-indexable and follow the hierarchy given in (3) (Öztürk and Pöchtrager, 2011).

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

Instead of representing the agreement paradigm in terms of theta-roles, we utilized the already proposed morphological features in the UD version of the Basque Dependency Treebank (Aranzabe et al., 2015). They mark the morphological case of the controller of the agreement as a language specific suffix to the features `Number` and `Person`. We adopted the same approach and used the following features: `Number[erg]={Sing,Plur}`, `Number[nom]={Sing,Plur}`, `Number[dat]={Sing,Plur}`, `Person[erg]={1,2,3}`, `Person[nom]={1,2,3}`, and `Person[dat]={1,2,3}`. This enables us to cover another language specific feature of Laz. The subject of an intransitive verb is marked with either the ergative or the accusative case according to the type of the verb (unergative or unaccusative) (Öztürk and Pöchtrager, 2011). By using this method for the agreement paradigm, we also mark the intransitive verb types indirectly. Unergative verbs which only have agent argument will be represented with `[erg]` layer and unaccusative ones with the `[nom]` layer.

3.3 Affirmative Preverbs

The verb in Laz has a highly complex structure which can host 16 different slots for inflection (Öztürk and Pöchtrager, 2011).[4] Four of these slots are used by preverbal affixes. One such type of a preverbal affix is the affirmative preverb. The affirmative preverbs (*ko-, do-, menda-, o-*) conveys either habitual reading or certainty reading.

For the annotation process, it presents two challenges. The first challenge is related to its syncretic functions. Without any context information, it is impossible to differentiate between the habitual reading and the certainty reading. For this treebank, we followed the glosses presented in the linguistic works.

The second challenge stems from its certainty reading. It signals that the speaker knows that the event has *certainly* happened or will happen (Öztürk and Pöchtrager, 2011). The only morphological feature related with this reading is the `Evident` feature within the UD framework. However, affirmative preverbs do not convey the source information. For example, the sentence (4) may be uttered even when the speaker does not witness Ali's coming.

(4) *Ali* *ko-mo-xt'-u.*
Ali.NOM PV_{aff}-PV_{spat}-come-PST.3SG
'Ali certainly came.' (Öztürk and Pöchtrager, 2011)

One possibility is adding a new value to the feature as `Polarity=Aff`. However, the rest of the values within the polarity feature is not directly related to the phenomenon in Laz. Additionally, the

[3] https://github.com/UniversalDependencies/docs/issues/734

[4] Morphological slots of the verb are as follow: PV_{aff}-PV_{spat}-Person-Valency-Root-AUG-CAUS.INTR-CAUS-TR-CAUS.PRF-TS-IMPF-SBJV-Person-COND-PL-AUX

name *affirmative* comes from its complementary distribution with the negative marker, which is pointed out to be irrelevant by Öztürk and Pöchtrager (2011). Affirmative reading is still possible without the preverb. Instead, we introduced a new value *certain* to the feature aspect as `Aspect=Crt`. The feature aspect is also used with its other reading as `Aspect=Hab`.

4 Parsing with a Multilingual Parser

Since the size of the treebank is insufficient to train a dependency parser for the Laz language, we instead observed the parsing success of a multilingual parser on our treebank without using any resources on the Laz language in the training phase. We chose to use UDify for this task. UDify is a state-of-the-art multilingual multi-task model that can predict annotations for any treebank annotated in UD style. The UDify model is fine-tuned on multilingual BERT pretrained embeddings (Devlin et al., 2019) and can syntactically annotate sentences in any language without requiring any language-specific components. We want to benefit from an automatic annotator in the hope that it will ease the manual annotation task of additional Laz text.

We used a pretrained multilingual UDify model to parse our treebank. Since the Laz language does not have any NLP resources other than this treebank, such as pretrained word embeddings or a publicly available corpus,[5] the language is completely unknown to the UDify model. Moreover, none of the language resources that were used in training of the UDify and BERT models belong to South Caucasian language family, which includes the Laz language.

Treebank	**Token count**	**UAS**	**LAS**
Our Laz Treebank	2K	44.15	29.05
Akkadian-PISANDUB[6]	1K	27.65	4.54
Amharic-ATT (Seyoum et al., 2018)	5K	17.38	3.49
Cantonese-HK (Wong et al., 2017)	6K	46.82	32.01
Erzya-JR (Rueter and Tyers, 2018)	15K	31.90	16.38
Komi Zyrian-IKDP (Partanen et al., 2018)	1K	36.01	22.12
Komi Zyrian-Lattice (Partanen et al., 2018)	2K	28.85	12.99
Naija-NSC[7]	12K	45.75	32.16
Sanskrit-UFAL[8]	1K	40.21	18.56
Warlpiri-UFAL[9]	$<$ 1K	21.66	7.96
Yoruba-YTB (Ishola and Zeman, 2020)	2K	37.62	19.09

Table 1: Test results for our treebank and some of the low-resourced treebanks in (Kondratyuk and Straka, 2019). The UDify and BERT models have no training data for any of these treebanks. The token counts given in the second column are from the UD v2.3 versions of the treebanks.

In their paper, Kondratyuk and Straka (2019) stated the success of their UDify model on every available treebank in the Universal Dependencies v2.3 corpus. Like the Laz language, there are other languages that are unknown to UDify in this corpus, although some of them have close relatives in the training data. In Table 1, we give the UAS and LAS scores of the model on our treebank as well as the scores of the treebanks used in (Kondratyuk and Straka, 2019) that were not used in the training phase of the UDify model for comparison. All of the treebanks in Table 1 have 15K or less tokens.

In Table 1, we observe that, the highest two attachment scores are achieved on Naija-NSC and Cantonese-HK, which is somewhat expected because Naija is an English-based creole language and

[5] Only existing corpus is mentioned in the work of Önal and Tyers (2019); however, they did not publish this treebank.
[6] `https://universaldependencies.org/treebanks/akk_pisandub/index.html`
[7] `https://universaldependencies.org/treebanks/pcm_nsc/index.html`
[8] `https://universaldependencies.org/treebanks/sa_ufal/index.html`
[9] `https://universaldependencies.org/treebanks/wbp_ufal/index.html`

Cantonese belongs to the Chinese language family. Although there is not any language resource in the training data which shares the same language family with Laz language, our treebank has the third best attachment scores. The low average token count per sentence (approximately 4 tokens per sentence) in our treebank has an impact on these relatively high scores. However, the results suggest that we need to manually annotate more data from scratch before taking advantage of a dependency parser as a pre-annotator. We see that there is an immense need to improve the parsing scores of our treebank and we hope that the parsing accuracy of the Laz language will greatly benefit from the presence of training data as we continue to annotate more sentences manually in the Laz language.

5 Conclusion

This paper reports the development of the first Laz Treebank ever. Considering that there is no Laz corpus or treebank that precedes this work and no treebank from the South Caucasian language family, we believe our work will be an important contribution to the field and support the further development of typological studies that utilize the UD framework. In addition, our work is also a gold standard for both syntactic and morphological annotation which will help many future studies on both Laz and Georgian. Our treebank currently consists of 576 sentences extracted from linguistic research. This paper discusses three topics of Laz grammar which can contribute to the UD framework in the future. Additionally, we report dependency parsing results with a multilingual parser: UDify. The parsing accuracy of our treebank in a zero-shot learning setting are found to be similar to other low-resourced treebanks with no training data.

The main aim of our work is to create treebanks for the minority languages spoken in Anatolia, including Cappadocian Greek, Pontic Greek, Pomak, Ladino, and many others. We believe that these series of treebanks can promote future research in language contact studies as well as NLP studies in these languages. Additionally, we hope to contribute to the revitalization efforts for these languages, including Laz. In the near future, we will expand our treebank to include 1,500 sentences before annotating any other language.

References

Maria Jesus Aranzabe, Aitziber Atutxa, Kepa Bengoetxea, Arantza Diaz de Ilarraza, Iakes Goenaga, Koldo Gojenola, and Larraitz Uria. 2015. Automatic conversion of the Basque dependency treebank to universal dependencies. In *Proceedings of the fourteenth international workshop on treebanks an linguistic theories (TLT14)*, pages 233–241.

Hagen Blix. 2020. Spans in South Caucasian agreement. *Natural Language & Linguistic Theory*, May.

Ömer Demirok, Deniz Özyıldız, and Balkız Öztürk. 2019. Complementizers with attitude. In Maggie 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.

Ömer Demirok. 2013. Agree as a unidirectional operation: Evidence from Pazar Laz. Master's thesis, Boğaziçi University.

Ömer Demirok. 2014. The status of roots in event composition: Laz. *Lingue e linguaggio, Rivista semestrale*, (1/2014):83–102.

Ömer Demirok. 2018. A modal approach to dative subjects in Laz. In Sherry Hucklebridge and Max Nelson, editors, *NELS 48: Proceedings of the Forty-Eighth Annual Meeting of the North East Linguistic Society*. CreateSpace Independent Publishing Platform.

Ömer Demirok. 2020. Non-linear blocking of portmanteaus: a case study on Laz. Talk given at NanoLAB, Masaryk University, Brno.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Cheikh Bamba Dione. 2019. Developing universal dependencies for Wolof. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 12–23, Paris, France, August. Association for Computational Linguistics.

Kira Droganova and Daniel Zeman. 2019. Towards deep universal dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 144–152, Paris, France, August. Association for Computational Linguistics.

Betül Emgin. 2009. *Finiteness and complementation in Laz*. Ph.D. thesis, Boğaziçi University.

Belma Haznedar. 2018. The living Laz project: The current status of the Laz language and Laz-speaking communities in Turkey. Talk given in LINGDAY, Boğaziçi University, Turkey.

Olájídé Ishola and Daniel Zeman. 2020. Yorùbá dependency treebank (YTB). In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5178–5186.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China, November. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Atul Kr. Ojha and Daniel Zeman. 2020. Universal Dependency treebanks for low-resource Indian languages: The case of Bhojpuri. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 33–38, Marseille, France, May. European Language Resources Association (ELRA).

Esra Önal and Francis Tyers. 2019. Building a morphological analyser for Laz. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 869–877, Varna, Bulgaria, September. INCOMA Ltd.

Balkız Öztürk and Markus A. Pöchtrager. 2011. *Pazar Laz*. Lincom Europa München.

Balkız Öztürk and Eser Erguvanlı Taylan. 2013. Omnipresent little v in Pazar Laz. Talk given at Little v Workshop, University of Leiden.

Balkız Öztürk. 2016. Applicatives in Pazar Laz. Talk given at The South Caucasian Chalk Circle 3, Paris.

Balkız Öztürk. 2019. The loss of case system in Ardeshen Laz and its morphosyntactic consequences. *STUF - Language Typology and Universals*, 72(2):193 – 219.

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first Komi-Zyrian universal dependencies treebanks. In *Second Workshop on Universal Dependencies (UDW 2018), November 2018, Brussels, Belgium*, pages 126–132.

Jack Rueter and Francis Tyers. 2018. Towards an open-source universal-dependency treebank for Erzya. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 106–118.

Binyam Ephrem Seyoum, Yusuke Miyao, and Baye Yimam Mekonnen. 2018. Universal dependencies for Amharic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Guillaume Thomas. 2019. Universal dependencies for Mbyá Guaraní. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 70–77, Paris, France, August. Association for Computational Linguistics.

Tak-sum Wong, Kim Gerdes, Herman Leung, and John Lee. 2017. Quantitative comparative syntax on the Cantonese-Mandarin parallel dependency treebank. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), September 18-20, 2017, Università di Pisa, Italy*, number 139, pages 266–275. Linköping University Electronic Press.