Building a Morphological Analyser for Laz

Esra Önal
Cognitive Science,
Boğaziçi University
Istanbul, Turkey
esra.onal@boun.edu.tr

Francis M. Tyers
Department of Linguistics
Indiana University
Bloomington, United States
ftyers@iu.edu

Abstract

This study is an attempt to contribute to documentation and revitalization efforts of endangered Laz language, a member of South Caucasian language family mainly spoken on northeastern coastline of Turkey. It constitutes the first steps to create a general computational model for word form recognition and production for Laz by building a rule-based morphological analyser using Helsinki Finite-State Toolkit (HFST). The evaluation results show that the analyser has a 64.9% coverage over a corpus collected for this study with 111,365 tokens. We have also performed an error analysis on randomly selected 100 tokens from the corpus which are not covered by the analyser, and these results show that the errors mostly result from Turkish words in the corpus and missing stems in our lexicon.

1 Introduction

The Laz language, which is mainly spoken on the northeastern coastline of Turkey and also in some parts of Georgia has been recorded as a 'definitely endangered[1]' language in UNESCO Atlas of the World's Languages in Danger. It belongs to South Caucasian language family[2] with the number of speakers estimated to be between 130,000 and 150,000 according to UNESCO 2001 records and between 250,000 and 500,000 according to more recent studies (Haznedar, 2018). Until the 1920s it was a spoken language with only some written collection of Laz grammar and folklore studies. Later, İskender Tzitaşi became the pioneer in developing a writing system for Laz based on Latin alphabet and later 'Lazuri Alboni' (Laz Alphabet) and only after 1990s, the written texts started to come out as several associations were founded for the preservation of Laz language and culture. Now with all these efforts, Laz has been thought in public schools in Turkey as an elective language course since 2013 (Kavaklı, 2015; Haznedar, 2018).

There is not much research on lexicon and syntax of Laz and the first academic level research studies began by the end of 20th century. In 1999, the first dictionary for Laz (Turkish—Laz) was prepared and published by İsmail Bucaklişi and Hasan Uzunhasanoğlu. The following years, Bucaklişi also published the first Laz grammar book (Kavaklı, 2015) and has begun teaching Laz at Boğaziçi University in İstanbul as an elective course since 2011. The foundation of the *Lazika Publishing Collective* in 2011 has given rise to the publication of more than 70 books on Laz language and literature (Kavaklı, 2015).

Laz language is only one of many that faces the danger of extinction. By the end of this century, many will not survive with the decreasing number of the native speakers of such languages (Riza, 2008). This has alarmed not only native speakers of these languages but also research community to direct their attention for language documentation as well as preservation and revitalization studies for these languages (Ćavar et al., 2016; Gerstenberger et al., 2017). Bird (2009) calls out for a 'new kind of computational linguistics' in his paper that would protect this endangered invaluable cultural heritage by helping to accelerate these studies, and he ends his paper with these words 'Who knows, we may even postpone the day when these languages utter their last words.' which emphasizes the importance of each and every attempt to keep these languages alive.

Riza (2008) gives accounts on language diversity on the Internet by pointing to the fact that many en-

[1] UNESCO defines the degree of definitely endangered as the situation in which "children no longer learn the language as mother tongue in the home".

[2] The Southwest Caucasian language family consists of four languages: Svans, Mingrelians, Georgian and Laz.

Proceedings of Recent Advances in Natural Language Processing, pages 869–877,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_101

dangered languages lacks access to Information and Communication Technology and the representation of these languages on digital environment is rather low. Considering Laz resources online, there are some online dictionaries and only a couple of web sites that give information about the Laz language and culture, and many of them are mostly in Turkish or English. He suggests that regarding 'digital language divide' such small regional languages must be represented more by creating and using resources in digital format.

One of the drawbacks while working with these languages is clearly the small amount of data to begin with (written or spoken, annotated or non-annotated) (Riza, 2008). Current dominant computational methods and tools are mostly used on languages with large corpora, following a statistical approach to train their systems according to a relevant task. However, with little data at hand these methods may not present a good solution. Therefore, Gerstenberger et al. (2017) suggests a rule-based morpho-syntactic modelling for annotating small language data. On their study of Komi language, his results show by-far significant advantages of rule-based approaches for endangered languages by providing much more precise results in tagging as well as 'full- fledged grammatical description based on broad empirical evidence' and a future development for computer-assisted language learning systems.

In this study, the aim is to create a morphological analyzer using the Helsinki Finite State Toolkit (HFST) that will help to overcome manual annotation of a potential Laz corpus. Additionally, as Gerstenberger et al. (2017) suggest, these may later help developing programs to be able to facilitate learning of Laz, considering the increase of interest in Laz courses not only in secondary schools but also in universities such as Boğaziçi University and İstanbul Bilgi University (Haznedar, 2018). From spelling and grammar-checkers to machine translation systems and language learning materials, this small study will hopefully lead to further developments on Laz language in the field of Computational Linguistics.

The remainder of the paper is laid out as follows: in Section 2, the grammatical structure of Laz is discussed and later, in Section 3.1 and 3.2 the process of preparing a lexicon and corpus for Laz is described. The Section 4 gives and explains the details of the morphological analyzer and the usefulness of Flag diacritics for representing Laz verbal complex and for generating complex verbal word forms.

2 Laz

There are eight dialects of the Laz language, none of which is considered to be normative or 'standard'. Even though underlyingly the structure of these dialects is the same, they show lexical and morphological, as well as phonological differences. There are two main groups. The Western dialects (Gyulva), such as Pazar (Atina), Çamlıhemşin (Furthunaşi gamayona), Ardeşen (Arthaşeni) and the Eastern dialects (Yulva) as Fındıklı (Viʒe), Arhavi (Arkabi), Hopa (Xopa), Borçka-İçkale (Çxala).[3]

For the purposes of this initial study we have chosen to base the analyser on the Pazar dialect. The reasons for this are twofold: Firstly the Pazar dialect is less irregular in terms of verbal inflection and secondly a separate and well-documented grammar of Laz written in English is based only on this dialect.[4] Unfortunately, there is no study yet that would provide an analysis of Laz grammar to be treated as 'standard' (Haznedar, 2018).

2.1 Verbs

In terms of morphosyntactic alignment, Laz is an ergative–absolutive language. It marks the subject of unergative predicates and transitives with agentive/causer subjects with ergative case while the subject of unaccusative predicates and the direct object of transitive and ditransitive verbs are inflected with nominative case. These patterns are marked differently on the verbal complex, depending on their case markings which also indicate their argument types. The verb encodes person information both preverbally and postverbally as seen in Table 1 and Table 2 and 3.[5] While we can observe verbs agreeing with agent-like arguments and sole argu-

[3] We exclude the Sapanca dialect as the region in which it is spoken is further away from the other dialects. Speakers of the Sapanca dialect are considered to be migrated from Batum, Georgia to Sapanca, Turkey (Bucaklişi and Kojima, 2003)

[4] The main grammar book used for this study is Pazar Laz written by Öztürk and Pöchtrager (2011) which is based on courses given by İsmail Bucaklişi in Boğaziçi University and it is the most recent and only complete study on a dialect of Laz written in English which would enable us to define grammatical rules for the morphological analyser. The grammar book by René Lacroix (2009) was also referred to several times but since it is mostly based on Arhavi (Arkabi) dialect and written in French, we used it when we needed to look for more examples for certain structures and specific details, especially valency-related vowels on the verbal complex and verb classes.

[5] It should be noted that post-verbal person markers encode tense information as well.

ments in both positions at the same time, we can only observe theme-patient (of mono-transitive and ditransitive verbs) and dative marked recipient-goal or applied non-core argument agreement in preverbal position. Additionally, Laz applies a hierarchical selection rule among arguments while marking person in -2 pre-verbal position seen in Table 1. This can be represented as in (1), where D represents the dative-marked arguments in the structure and P represents theme/ patient argument type while A means agent-like argument type.[6]

(1) D1/2 > P1/2 > A1 > D3=P3=A2/3

The reason why D1/2 arguments comes first but not D3 is that D3 is unmarked; therefore, overt A/S1 markings fills the position if they are available in the structure. Only when we have A2/3 type argument, the position remains empty. We will also discuss this topic later in Section 4.1.

(2) *Bere-k Lazuri d-i-gur-am-s.*
child-ERG Laz.NOM PV-VAL-learn-TS-PRS.3.*a*.SG
'The child is learning Laz.'

(3) *Bere-s Lazuri dv-a-gur-e-n.*
child-DAT Laz.NOM PV-APPL-learn-TS-PRS.3.S.SG
'The child is able to learn Laz.'

The case markings of arguments are apt to change despite their argument type when the general construction of the predicate changes, which in turn changes the verbal complex as in present perfect constructions and in expressing ability and involuntary actions.[7] They lead to the backgrounding of agent-like arguments; therefore, we can only observe 3.SG in post-verbal person marking position unless NOM marking objects are emphasized. Emphasizing such arguments allow the verbal complex to bear their marking in post-verbal position, as in (2) and (3) from Öztürk and Pöchtrager (2011).

As seen in the example, the verb not only takes *a-* valency vowel and ability related *-e(r)* TS suffix but also changes the person marking as well. There are also valency-changing operations such as applicativization, causativization and reflexivization which introduces non-core dative marked arguments, nominative and dative-marked arguments, and verbal reflexivization through a theme argument or a non-core argument[8] respectively. All these operations commonly mark the verb with different valency-related vowel in the same preverbal position. Therefore, under conditions where the verb is needed to be inflected both applicativization and causativization at the same time, APPL vowel `i/u-` suppresses CAUS vowel `o-`. This is important for us since when we mark the verb with APPL, the structure should allow CAUS construction as well. We will discuss such intersecting constructions and how we deal with them in our lexicon file in Section 4.1 in detail. An example of this from Öztürk and Pöchtrager (2011) is given in (4).

(4) *Him Ayşe-s bere*
S/he Ayşe.DAT child-NOM
u-bgar-ap-ap-u-n
APPL-cry-CAUS-CAUS.PERF-TS-PRS.3.A.SG
'S/he has made Ayşe make the baby cry.'

Table 1 shows the pre-verbal complex of Pazar Laz and Table 2 and 3 show post-verbal complex which we have based our main FST continuation classes on.

2.2 Substantives

Adjectives, adverbs and nouns together constitute substantive category in the language since they behave similarly within a sentence depending on the suffixes they carry as well as their position. An adverb can take dative suffix *-s* and an adjective can be used as a noun by taking case or plural marker. The differentiation between these categories are not very clear.

As mentioned partially above, Laz marks nouns with case markings such as ergative *-k*, nominative (unmarked), dative *-s*, allative *-şe* (showing the direction of an event), ablative *-şe(n)*[9] (indicating the source of an event), genitive *-şi* and instrumental -te. Other than these case markings, nouns are marked plural marking *-pe* and only some nouns ending with *a* take *-lepe* as plural marker. Since there is no phonological rule for this alternation, we need to categorise each noun in our lexicon manually for our morphological analyser.

2.3 Orthography

Orthographically, we have adopted the Laz alphabet given below which is an extended version of Turk-

[6] Intransitive verbs only has S 'sole' argument which has the same marking pattern with A arguments.

[7] These three constructions are called *inversion* constructions which require certain type of predicates, specifically those including agent-like arguments, and ergative case is never available for these constructions.

[8] This additionally assumes the function of applicativization.

[9] Final *n* only occurs with post-position *doni*; therefore, we will mark every noun with both forms.

-4	-3	-2	-1	0
Affirmative preverb	Spatial preverb	Person marker	Valency-related vowel	**Root**
o-, ko-, do-, menda-	ama-, ce-, cela-, ç̌eǩo-, ç̌eşǩa-, do-, dolo-, e-, eǩo-, ela-, eşǩa-, eǯo-, eyo-, gama-, go-, gola-, goyo-, ǩoǩo-, ǩoşǩa-, me-, mela-, menda-, meşǩa-, meyo-, mo-, mola-, moǩo-, moşǩa-, moǯo-, moyo-, oǩo-, exo-, ǩoǯo-, oxo-, gela-, ǩoǯa-	*S-A.1: v-,p-, p̌-, b, (f-) **P-D.1: m- P-D.2: g-, k-, ǩ *‘S-A’ = Sole or agent-like arg **‘P-D’ = Patient/theme arg or Dative marked arg	i/u-, i-, a-, o-	-t̆ax- ‘break’

Table 1: Pre-verbal complex the numbers in the header refers to the pre-verbal position relative to the verbal root. The spatial preverb is a prefix that indicates the direction or manner of an event. The different forms of person markers are realised based on the laryngeal properties of the following consonant. This will be later discussed in Section 4.2.

0	1	2	3	4	5	6
Root	Augment. stem formant	Causative suffix for intransitives	Causative suffix for transitives	Cusative suffix for present perfect cons.	Thematic suffix	Imperfect stem formant
-t̆ax- ’break’	-am	-in	-ap	-ap	-am,-um,-er,-ur	-ť

Table 2: Post-verbal complex-1

7	8	9	10	11
Subjunctive marker	Person suffixes	Conditional marker	Plurality	Auxiliaries
-a	S-A.1.PST: -i S-A.2.PST: -i S-A.3.PST: -u S-A.1.PRS: ∅	-ǩo	-t; -es, an (3rd)	-(e)re, -(e)rťu

Table 3: Post-verbal complex-2

ish alphabet based on Latin letters. Across texts, they have been encoded with different characters but these forms will be the standard for our study.

3 Resources

3.1 Lexicon

The lexicon composed for this study comes from the *Büyük Lazca Sözlük* (Didi Lazuri Nenapuna). It is the most extensive dictionary available for Laz prepared by Hasan Uzunhasanoğlu , İsmail Bucaklişi and İrfan Çağatay Aleksiva in 2007 in Laz and Turkish.

The verbs were extracted from the dictionary automatically whereas other word classes were extracted semi-automatically. The words are taken as entries with their dialect labels[10] and if available, dialect-specific forms as seen in (5).[11]

(5) doinu [Atn., Viw, dorinu Gyl., Ark., Xop., doǩunapa Sap.]

We have prepared verb word lists for each dialect separately as well as a complete word list for all. Considering the possibility that dialects may borrow words from one another, we decided to build a lexicon based on not only the Pazar dialect but all dialects of Laz. This is an important strategy to form a 'common source lexicon' (Beesley and Karttunen, 2003). However, for the sake of simplicity, we have excluded nominal and verbal compounds from our lexicon.

The challenging part in preparing the lexicon has been the stemming process for verbs since the verbs in the dictionary are in their infinitival form and some of them also include preverbs. Even though the preverbs have been easily separated, the infinitive suffixes were harder to process. For example, there are verbs ending with *-alu* and while some of these verbs include *-al* suffix in their bare form, some do not. It means that they are lexically determined.

Even though noun declension was easy to define, extracting substantives from the dictionary and carefully separating them into nouns, adverbs, and adjectives as well as categorizing other syntactic elements like interjections, conjunctions and pronominals were among the hardest tasks for this study. We needed to separate these word classes semi-automatically because there were words that should be put in more than one category such as in both noun and adjective or adjective and adverb (determined only by sentential position) or noun and adverb. Therefore, it could not be possible for us to include words (except verbs) that belong to other dialects other than Pazar for this study.

3.2 Corpus

We have collected different type of written texts for our Laz corpus. However, differences in terms of dialects have forced us to divide texts into their corresponding dialects for this study since we have decided on working Pazar Laz first the reasons of which are discussed in Section 2. Unfortunately, Pazar Laz has almost no written text known in the literature. The only resource we have is an 800 page document consisting of 111,365 tokens collected by İsmail Bucaklişi, a native speaker of Pazar Laz, by himself which contains daily conversations and stories shared in his immediate circle. It should be noted that it also contains Turkish words and sentences given as translations throughout the document the effects of which on the results can be seen in Section 5.2.

4 Methodology

The purpose of this project is to develop a computational model for morphological analysis for Laz by using the Helsinki Finite-State Toolkit (HFST; (Linden et al., 2011)) which is popular in this field of research. A finite-state transducer associates a morphological analysis with the corresponding phonological representation. Xerox `lexc` and `twolc` formalism supported by HFST are used to create lexicon files and a two-level grammar file respectively (Beesley and Karttunen, 2003).

4.1 Lexicon Files

The `lexc` (**Lex**icon **C**ompiler) formalism is used to define lexicons which contain grammatical labels and morphotactic rules for the morphemes in the language (Beesley and Karttunen, 2003).

4.1.1 `lexc` File for Substantives

The substantive `lexc` file has 27 tags for morphemes indicating person, number and case information and 18 continuation classes for morphotactics or word-formation rules together with the

[10] The following dialect codes were found in the dictionary: *'Yul' (Eastern dialects), 'Gyl' (Western dialects), 'Viw' (Viʒe), 'Xop' (Xopa), 'Ark' (Arkabi), 'Çxl' (Çxala), 'Atn' (Atina/Pazar), 'Fur' (Furthunaşi gamayona), 'Arş' (Arthaşeni), 'Sap' (Sapanci).*

[11] (5) shows that *doinu* 'to give birth' belongs to *Atn.* and *Viw.* dialects, and it takes the form of *dorinu* in *Gyl.*, *Ark.* and *Xop.* dialects, and *doǩunapa* in *Sap.* dialect.

a	*b*	*c*	*ç*	*ç̌*	*d*	*e*	*f*	*g*	*ğ*	*h*	*x*	*i*	*j*	*k*	*ǩ*	*l*
[a]	[b]	[dʒ]	[t͡ʃ^{h}]	[tʃ’]	[d]	[e]	[f]	[g]	[ɣ]	[h]	[x]	[i]	[ʒ]	[k^{h}]	[k’]	[l]
m	*n*	*o*	*p*	*p̌*	*r*	*s*	*ş*	*t*	*ť*	*u*	*v*	*y*	*z*	*ž*	*ʒ*	*ǯ*
[m]	[n]/[ŋ]	[p]	[p^{h}]	[p’]	[r]	[s]	[ʃ]	[t^{h}]	[t’]	[u]	[v]/[w]	[j]	[z]	[dz]	[ts^{h}]	[t͡s’]

Table 4: The 34 letters of the Laz alphabet. The letters ǩ p̌ ť ç̌ ǯ ʒ and ž represent ejective consonants.

lexemes. We specified pronouns as personal pronouns, possessive pronouns, demonstratives, reflexives, interrogative pronouns, indefinite pronouns, quantifiers as well as numerals in different continuation classes. We have continuation classes for case and plural markers to show nominal inflection. There are two forms of plural markings and ablative markings each, whose differentiation is lexical, not phonological. Therefore, we have encoded this information in our lexicon by using flag diacratics that will be explained in Section 4.1.3.

4.1.2 `lexc` File for Verbs

The `lexc` file for Laz verbal complex has 53 tags for the morphemes encoding preverb, valency-related, mood, tense, person and number information and 19 continuation classes which correspond to the affixes in the verbal complex as defined in Öztürk and Pöchtrager (2011) also seen in Table 1, 2 and 3 with three additions — additive position for suffix *-ti*, — question for *-i*, — participle for *-eri*.

We have mostly followed the description in Öztürk and Pöchtrager (2011) when naming the tags and classes.

The final combined `lexc` file also includes interjection, conjunction, negation, post-position and pre-position lexicons with 5 more tags.

4.1.3 Flag Diacritics

Laz verb complex has required substantial use of *flag diacritics*[12] to solve problems like dependent person marking, and causativisation or applicativisation processes, which require preverbal valency-related vowel marking as well as postverbal causative markers at the same time. The hierarchical selection rule for person marking position preverbally among the arguments of the verb is easily applied using flag diacritics. We have allowed structures with 3rd person prefixes to only occur in structures with 3rd person suffixes by disallowing paths including 1st and 2nd person prefixes. This is done by setting a flag, @P.D-P.3@ which means **P**ositive setting of the D-P (dative–patient) type argument bearing 3rd person information and later in person suffix continuation class we reject/**D**isallow those paths with positively setting of 3rd person information by setting @D.D-P.3@ for the 1st and 2nd person suffixation. Additionally, we can also reject paths that include combinations of 1st and 2nd person D-P with 1st and 2nd person A (agent) respectively since the language does not allow such constructions. We use the same patterns as above for these rules.

We have also found flag diacritics useful for overwriting of valency-related vowels. In such constructions, we engage in two separate operations/constructions at the same time such as causativisation and present perfect construction both of which mark the verb with their specific valency-related vowel in the -1 position. However, Laz allows overwriting causative *o-* to be overwritten by applicative *i-* while keeping post verbal causative markers *-in* or *-ap*. We have managed to form these constructions by also allowing applicative *i-* (as well as causative *o-*) to have flag @P.CAUS.PRS@ which will let them through paths defined with @R.CAUS.PRS@ (**R**equire the causative feature to be present). These paths are naturally those causative suffixes which do not allow structures with related valency vowel otherwise.[13]

Other flag diacritics include N which sets the related feature as **N**egative. In our study, we use them for subjunctive suffix and its special construction with thematic suffixes. They do not normally occur together but we can see that they do in constructions with the imperfective suffix in between in (6) from Öztürk and Pöchtrager (2011).

[12] Flag diacritics are used for feature-setting and feature-unification operations. They represent long-distance constraints for dependencies within a word (Beesley and Karttunen, 2003). As a member of the same language family, Georgian also shows these kinds of long-distance dependencies in verbal complex which are effectively treated again with this device in order to build a computational grammar for Georgian (Meurer, 2009).

[13] Additionally, Laz allows only intransitive bases to take the *-in* causative suffix, so we have also tried to use flag diacritics to differentiate between transitive and intransitive bases by encoding the information onto verb itself. However, since we have automatically extracted verbs from the dictionary, we could not label all of them (2240 verb roots) as transitive or intansitive for this study, we ignore this differentiation and allow all verb roots to be able to bear both *-in* and *-ap* suffixes.

(6) *m-i-t́ax-ap-ur-t́-a-s*
D1-APPL-break-CAUS.PERF-TS-IMPRF-SUBJ-PRS.3.S.SG
'Let say that I have broken it.'

If the subjunctive follows imperfective (sets thematic suffix information as @N.THM.PRS@), the path allows subjunctive (normally disallows thematic suffixes as @D.THM.PRS@) to follow thematic suffix after imperfective; therefore, we need to get rid of @P.THM.PRS@ setting by re-setting the same feature to N as @N.THM.PRS@ which will allow the structure to go through the path by taking non-past person markers that are set as @D.TNS.PST@ disallowing past tense constructions.

The substantive `lexc` file includes only one flag diacratic case which is to label nouns that take *-lepe* plural marker instead of *-pe*. We label the noun root with @P.LEPE.PRS@ in order for it to be able to take the path with @R.LEPE.PRS@ label.

4.2 `twol` File

The `twolc` (**Two-l**evel **C**ompiler) formalism is used to define phonological and morphophonological alternations. The `twol` file mostly includes person marking elements differing based on the following consonant's laryngeal property for verbal inflection. We define rules/environments to account for morphophonological changes in the structure with archiphonemes[14] as given in Figure 1.

Laz also exhibits a phonological change in noun stems starting with *n* sound when preceded by ejective *p̌*-, the person prefix for 1.SG. The two consonants are combined and becomes *m*. We represent this as ejective *p̌* turning to *m* and dropping the initial *n* of the stem. Additionally, the final *i* sound of noun stems becomes *e* when the stem is inflected with plural marker.

```
"Assimilation of person prefix to p-"
{V}:p <=> _ >: Voiceless: ;

"Assimilation of person prefix to b-"
{V}:ṕ <=> _ >: Ejectives: ;
```

Figure 1: Two two-level phonological rules for assimilation. The underspecified prefix archiphoneme {V} is restricted to surface either as *p* before voiceless consonants or *ṕ* before ejective consonants.

We also observe a morphologically-conditioned phonological alternation for valency-related vowels. The alternation for valency-related vowels *i/u-* depends on the preverbal person information, *i-* for 1st and 2nd person, and *-u* for 3rd person.

The preverbs show a great amount of morphophonological alternations in their final vowels, such as *a, e* and *o*. When they combine with overt person prefixes (consonants) together with valency related vowels, final *o* and *a* become *e* or *o* and the change is not always predictable. They can also turn into *v* or can be dropped. Even though they may end with the same vowel, the alternations can be different when followed by the same sound; therefore; we need to define different archiphonemes for the same vowel. For example, the final *o* sound in *exo-* drops when it attaches to a verbal complex starting with *a* sound but not the one in *moyo-*.

5 Results

We have evaluated the morphological analyser by calculating the naïve coverage and doing error analysis on randomly selected 100 tokens from the corpus.

5.1 Coverage

The coverage is measured by calculating the number of the tokens that receive at least one morphological analysis by the analyser. It should also be noted that the tokens may have other analysis that is correct but not provided by the analyser even though they get at least one analysis.[15] We have collected a corpus for Pazar Laz which consists of 111,365 tokens mentioned before in Section 3.2. The final morphological analyser has 64.9% coverage over this corpus.

Corpus	Tokens	Coverage
Pazar Laz	111,365	64.9%

Table 5: Naïve coverage of the analyser

5.2 Error Analysis

We have looked at the tokens that are not covered by the morphological analyser. Randomly selected 100 tokens has been examined and separated according to their error type seen in Table 7. It should also be

[14] An archiphoneme is used as a placeholder to be later replaced with the appropriate sound determined by morphophonological rules written in `twol` file. They are given inside curly brackets.

[15] Unfortunately since there is no annotated corpus which can be used as the 'gold standard', we were unable to calculate *precision* and *recall* that could show the average accuracy of the analysis provided by the transducer.

Category	Number of Stems
Noun	9417
Verb	2240
Adjective	745
Adverb	215
Pronoun	92
Numeral	46
Interjection	31
Postposition	29
Conjunction	8
Preposition	3
Negation	4
Total	12,830

Table 6: Number of lexicon entries by part of speech / lexical category.

noted that some of them may go into more than one category.

The highest percentage of unrecognized tokens belongs to the category of 'Missing lexeme'. This is partly because of the fact that our lexicon for substantives was not large enough to account for the phonological and lexical differences for stems in different dialects. For example, *açkvaneri* 'next time' appearing in our corpus belongs to Xopa dialect but not to Pazar dialect according the the dictionary. It also includes lexemes which do not appear in the dictionary and consequently not in our lexicon as well as those which are simply missed out during the automatic extraction of words from the dictionary.

We still have certain morphotactic rules to work on to be able to cover inflectional morphology of Laz such as verb inflection for adverbial clauses such as the *-şa* suffix meaning 'while' (related to the allative suffix normally attached to nouns).

6 Future Work

Since we still have problems with verb stemming and separating substantives into nouns, adjectives and adverbs as well as determining other word classes and their inflectional morphology, our current lexicon can be manually checked and extended accordingly. We definitely need to improve the coverage of the morphological analyser not only for Pazar Laz but also for other dialects of Laz for the future studies. This requires both defining morphotactics for other dialects and carefully separating and including lexemes from the dictionary. It is also equally important to prepare a gold standard corpus for Laz to be able to evaluate the accuracy of the analyser. Additionally, investigating borrowed words from specifically Turkish and how they are adapted and used in Laz will also improve the results. Derivational morphology is also nontrivial to look into to expand the lexicon.

Error Type	Frequency	Percentage
Missing lexeme	41	37.9%
Turkish word	35	32.4%
Missing or erroneous morphotactic rule	13	12.0%
Typing errors	7	6.4%
Loanwords	7	6.4%
Missing or erroneous Phonological rule	5	4.6%
Total	108	

Table 7: Error analysis for randomly selected 100 tokens

7 Concluding Remarks

We have presented the first ever morphological analyser for Laz, a language in the Caucasian language family spoken in Turkey. The analyser currently covers the Pazar dialect.

This study will hopefully lead to further studies for language documentation and revitalization efforts for Laz in a larger context.

All the up-to-date project files have been uploaded on `Github`[16] and licensed under the CreativeCommons BY-NC-SA 3.0.

8 Acknowledgements

We would like to express our sincere gratitude to İsmail Bucaklişi for his enormous contribution to the project as our native speaker informant as well as his help in finding language resources.

[16] https://github.iu.edu/esraonal/laz-morphological-analyser-fst

References

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite state morphology*. CSLI Publications.

Steven Bird. 2009. Natural language processing and linguistic fieldwork. *Computational Linguistics* 35(3):469–474.

İsmail Bucaklişi and Goichi Kojima. 2003. *Laz Grammar (Lazuri Grameri)*. Chiviyazilari.

Ciprian Gerstenberger, Niko Partanen, and Michael Rießler. 2017. Instant annotations in ELAN corpora of spoken and written Komi, an endangered language of the Barents Sea region. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Association for Computational Linguistics, pages 57–66.

Belma Haznedar. 2018. The living Laz project: The current status of the Laz language and Laz-speaking communities in Turkey.

Nurdan Kavaklı. 2015. Novus Ortus: The awakening of Laz language in Turkey. *İdil Journal of Art and Language* 4(16):133–146.

René Lacroix. 2009. *Description du dialecte laze d'Arhavi (caucasique du sud, Turquie) Grammaire et textes*. Ph.D. thesis, Université Lumière Lyon.

Krister Linden, Miikka Silfverberg, Erik Axelson, Sam Hardwick, and Tommi Pirinen. 2011. *HFST–Framework for Compiling and Applying Morphologies*, volume 100 of *Communications in Computer and Information Science*, pages 67–85.

Paul Meurer. 2009. A computational grammar for georgian. In Peter Bosch, David Gabelaia, and Jérôme Lang, editors, *Logic, Language, and Computation*. Springer Berlin Heidelberg, Berlin, Heidelberg, pages 1–15.

Hammam Riza. 2008. Indigenous languages of Indonesia: Creating language resources for language preservation. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Balkız Öztürk and Markus A. Pöchtrager. 2011. *Pazar Laz*. LINCOM.

Malgorzata Ćavar, Damir Ćavar, and Hilaria Cruz. 2016. Endangered language documentation: Bootstrapping a Chatino speech corpus, forced aligner, asr. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).