MORPHLAZ: A FINITE-STATE MORPHOLOGICAL ANALYZER FOR LAZ

ESRA ÖNAL

BOĞAZİÇİ UNIVERSITY

2021

MORPHLAZ: A FINITE-STATE MORPHOLOGICAL ANALYZER FOR LAZ

Thesis submitted to the
Institute for Graduate Studies in Social Sciences
in partial fulfillment of the requirements for the degree of

Master of Arts
in
Cognitive Science

by
Esra Önal

Boğaziçi University
2021

DECLARATION OF ORIGINALITY

I, Esra Önal, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;

- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;

- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature: ______________________________

Date: ______________________________

ABSTRACT

MorphLaz: A Finite-State Morphological Analyzer for Laz

This thesis is a part of documentation and revitalization efforts of the endangered Laz language, a member of South Caucasian language family mainly spoken on the northeastern coastline of Turkey. It introduces the implementation of the first automatic language analysis tool for Laz, specifically for Pazar dialect designed as a rule-based morphological analyzer developed with two-level morphology using finite-state networks. Additional language resources such as lexicon and corpus were collected for the purposes of increasing the coverage power and evaluating the performance of the analyzer.

Morphologically rich languages create many challenges for natural language processing (NLP) tasks. In order to develop high or low-level NLP systems such as lemmatization, part-of-speech-tagging, spelling correction and machine translation, in any NLP pipeline, the first aim is usually to do some sort of morphological analysis on text or speech. Among different approaches to the computational study of morphology, for this study, due to the low amount of language and computational resources, I chose a rule-based approach that is highly accepted and used for formalizing morphotactics and morphophonemics, namely two-level morphology and finite-state transducers.

The evaluation is based on naïve coverage of the analyzer over text data and error analysis. The results show 78.2% of coverage over the unique tokens in Pazar Laz corpus (PLC), 92.1% of coverage over Laz Treebank and 74.3% on Fındıklı Laz corpus (FLC). Error analysis on PLC results indicates that most of the word forms that could not be analyzed are due to missing word stems.

ÖZET

MorphLaz: Laz için Sonlu Durum Biçimbilimsel Çözümleyici

Bu tez, ağırlıklı olarak Türkiye'nin kuzeydoğu kıyı şeridinde konuşulan ve Güney Kafkas dil ailesi üyesi nesli tükenmekte bir dil olan Lazca'nın, hesaplamalı dilbilim perspektifinden belgelenmesi ve yeniden canlandırılması çalışmalarının bir parçasıdır. Sonlu durum teknolojisi ve iki seviyeli morfoloji kullanılarak Lazca'nın Pazar lehçesi üzerine geliştirilen, kural tabanlı bir morfolojik çözümleyici olarak tasarlanan ilk otomatik dil analiz aracının uygulamasını sunar. Sırasıyla kapsam gücünü artırmak ve çözümleyicinin performansını değerlendirmek amacıyla sözlük ve derlem gibi ek dil kaynakları toplanmıştır.

Herhangi bir ardışık işleme yapan NLP boru hattında kök çözümleme, sözcük türü etiketleme, yazım hataları düzeltme ve makine çevirisi gibi yüksek veya düşük seviyeli NLP sistemleri geliştirmek için, ilk amaç genellikle metin veya konuşma üzerinde bir tür biçimbilim analizi yapmaktır. Biçimbilim hesaplamalı çalışmasına yönelik farklı yaklaşımlar arasında, bu çalışma için, dil ve hesaplama kaynaklarının azlığı nedeniyle, biçim bilgisi ve biçimbilimsel ses bilgisini tanımlamak için yüksek oranda kabul gören kural tabanlı bir yaklaşımla iki düzeyli biçimbilimi ve sonlu durum dönüştürücüleri kullandım.

Değerlendirme, metin verileri üzerinde çözümleyicinin naïve kapsamına ve hata analizine dayanmaktadır. Sonuçlar, çözümleyicinin Pazar derleminde bulunan özgün kelimelerin %78.2'sini, Laz Treebank'in %92.1'ini ve Fındıklı lehçesi derleminin (FLC) %74.3'ü üzerinde kapsamı olduğunu göstermektedir. PLC sonuçlarındaki hata analizi, analiz edilmeyen kelime biçimlerinin çoğunun eksik kelime köklerinden kaynaklandığını göstermektedir.

ACKNOWLEDGEMENTS

I want to talk about people who have done and sacrificed as much as me and in fact maybe even more than me for my M.A. studies leading up till now to writing and finishing this thesis with these last words.

First, I would like to thank my thesis advisors, Dear Assoc. Prof. Arzucan Özgür and Prof. Balkız Öztürk for their unconditional support and patience for my studies and thesis during my very long 4 year M.A. journey although I was not very present at times. If Arzucan Hoca did not accept my consent request for an already very much over-quota introductory computer programming course during the last semester of my undergraduate studies, probably I would not be here where I am now, being almost finished with my M.A., accepted to a Ph.D. program in Linguistics (which also would not be possible without her) and continue doing what I am most passionate about, working on computational linguistics and trying to understand and formalize the mechanism behind languages. Again with her support, I was able to join an exchange program in the Linguistics Department at Indiana University in the USA where I actually started working on my thesis and Laz language.

Although I did not know Balkız Hoca or take her course before I asked her if she would accept me to be her student, she has been very supportive. With her help and the resources and funding from UYGAR Dil Merkezi under the Linguistics Department at Boğaziçi, I was able to present my project on Laz in two different conferences abroad. I also would like to thank her for her substantial research and academic efforts on Laz language which were referenced abundantly in this thesis justifiably since they were a great guide in understanding the complexities of Laz and in developing an automatic morphological analyzer.

I cannot thank enough my current Ph.D. advisor and mentor Assist. Prof Francis M. Tyers at Indiana University without whom this thesis would not exist. His

feedback was the most valuable for making this project happen.

My dear M.A. thesis committee member and professor from the Department of Computer Engineering, Prof. Tunga Güngör is one of the kindest professor I have ever had, who is the first to teach me about natural language processing and computational linguistics. I consider myself the luckiest to even have a chance to be in his class.

I also thank my other M.A. thesis committee members Dear Assist. Prof. Ümit Atlamaz and Dear Prof. Olcay Taner Yıldız for their support and feedback with their knowledge and extensive experience in the field. Their feedback shaped the result of this thesis in the most substantial ways and shaped my ideas for further studies.

One of the reasons I decided to do my M.A. at Boğaziçi University was my professors in the Department of Linguistics because I have always known and trusted that I could learn from them so much with their being the best in their field of research. Therefore, I also thank all and Prof. Sumru Özsoy, Prof. Aslı Göksel, Assoc. Prof. Elena Guerzoni and Assist. Prof. Pavel Logačev whose courses shaped and enriched my scientific eye and research in linguistics.

This section would not be named acknowledgements if I don't thank my undergraduate advisor from Foreign Language Education Department at Boğaziçi, Aylin Ünaldı since she is the first person who introduced me to this program and encouraged my passion on computational linguistics. She believed in me when I didn't and now, I think she proved me wrong. I will always be in her dept and be grateful for her support.

I would like to thank İsmail Bucak'lişi for his most generous contribution and support to any project I have done on Laz. His efforts to revitalize Laz are so extraordinary that they have encouraged me to do my absolute best.

Although we had very little, my family never stopped supporting me both emotionally and financially during my formal education. Since I was very little, my father, Yahya Önal and my mother, Fatma Önal have believed in me and guided me so that I could finish university and be an independent woman in a region where most woman are not encouraged to continue their education. I am, therefore, forever grateful for their unprecedented love and support, and I would like to dedicate this thesis to them. Also, my sisters, Zehra Yıldırım and Nazife Özer have always been there when I was overwhelmed and needed motivation. They would always spare one room in their home just so that I can come and study anytime.

My lifelong friends, Elif Yurt and Şeyma Kılınç have never left my side when I have felt most desperate since my undergraduate years even though we were not always in the same city or even in the same country.

My very (and very) dear Boğaziçi family Betül, Dilara, Görgüner, Selcen, Esra, Cumhur, Oğuz, Faruk, Erdem and Gültekin has thought me the real meaning of friendship. Every one of them helped me so many times and without any expectations during very stressful times that I don’t think I can ever pay them back. They have always encouraged me and made me believe that I can do anything even though I feel the absolute opposite. They are the ones who always bring me back to life when I lose sense of time studying and remind me to enjoy life to the fullest.

Fınally, I would like to thank TUBITAK for funding me during my M.A. degree as part of 2210/A National MSc/MA Scholarship Program which allowed me to focus on my studies more and contribute to the world of science.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Aim

In this thesis, the aim is to implement a rule-based morphological analyzer and generator for Laz language, particularly for the Pazar/Atina dialect using finite-state techniques, and additionally to collect written Laz corpora from personal accounts and literary books some of which are translated from other languages. The resources to be presented in this thesis constitute the basic components of many computational linguistics (CL) and natural language processing (NLP) tasks and applications.

1.2 Language and technology

"Who knows, we may even postpone the day when these languages utter their last words." (Bird, 2009, p. 473)

Laz language is only one of many that face the danger of extinction. By the end of this century, many will not survive due to the decreasing number of speakers (Hammam, 2008). This has alarmed not only native speakers of these languages but also researchers to join language preservation and revitalization efforts (Ćavar, Ćavar, & Cruz, 2016; Gerstenberger, Partanen, & Rießler, 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. Therefore, it is important to introduce these preliminary computational materials and tools in order to draw more attention to Laz language in the field of Human Language Technology.

Bird (2009) suggests that there must be new research collaborations between computational linguists and field linguists that will include collecting and analyzing data from endangered languages and contributing together to the development of necessary tools, methods and resources to facilitate the preservation or revitalization of these languages. He particularly defines a system that will flourish collaborative creation of a database of language data, later being annotated and ‘continuously expandable and fully exportable for local processing’.

Hammam (2008) gives accounts on language diversity on the Internet by pointing to the fact that many endangered languages lack access to Information and Communication Technology and the representation of these languages on the digital environment is rather low. Although there are only some online dictionaries and web sites that give information about Laz language and culture, many of them are mostly in Turkish or English. He suggests that regarding ‘the digital language divide’ such small regional languages must be represented more by creating and using resources in digital format.

The research on under-represented and low-resource languages in computational linguistics starts with the creation of language resources (Hammam, 2008) so that they can be used in developing computational tools that would help to facilitate the effort for language documentation (Ćavar et al., 2016; Gerstenberger et al., 2017) as well as increase the use of such languages among speakers.

1.3 The organization of the thesis

This thesis is laid out as follows: Chapter 2 is an introduction to the computational study of morphology and how finite-state technologies and two-level morphology are used for morphological analysis and synthesis. Chapter 3 provides an extensive overview of Laz language and grammar towards understanding the morphotactics and morphophonemics of the language which constitute the two main components of the analyzer. In addition, dialectical variations are discussed in detail in this chapter to justify certain decisions and strategies while developing the analyzer such as creating a ‘common source lexicon’ for Laz. In Chapter 4, I discuss the implementation of the morphological analyzer as an FST and introduce Xerox *lexc* and *twolc* formalism in which morphotactic and morphophonemic rules are defined, and HFST command line tools to compile these rules into transducers. Finally, in Chapter 5, I first explain the process of collecting Laz corpora and creating the lexicon for the analyzer, and then, present the results on the performance of the morphological analyzer on three different corpora. I conclude the thesis by talking about the challenges I have come across during this study and how to extend this morphological analyzer to include other dialects in the final chapter.

CHAPTER 2

COMPUTATIONAL STUDY OF MORPHOLOGY

2.1 Introduction

"*Knowledge of a language includes knowledge of the systematicity in the relationship between the form and meaning of words.*" (Booij, 2012, p. 4)

As the field of morphology is basically about the study of word structure and word formation in natural languages, the main task in computational morphology is to provide a computational analysis or synthesis of word forms or in other words, to take a morphologically complex word as input and provide an analysis of the components, namely *morphemes* and/or their morphosyntactic properties as output. The analysis in 1 shows three types of outputs which can be potentially provided with stemming/lemmatization[1], segmentation, and morphological analysis for the word form *k’at’upe* ‘cats’ in Laz.

(1) *k’at’upe ‘cats’* →*k’at’u* (morpheme; stem/lemma)
→*k’at’u* + *pe* (+ morpheme; suffix)
→*k’at’u*<*noun*> + <*plural*> (+ morpho. features/tags)

In this chapter, I will talk about the importance of computational analysis of morphology for NLP research in Section 2.2, and then introduce and compare different approaches to automatic morphological analysis in Section 2.3. Finally, Section 2.3.1 and 2.3.2 will be specifically about finite-state technologies and two-level morphology for creating computational morphologies.

2.2 Importance of computational morphology

Computational study of morphology is important in the fields of CL and NLP for any language, but especially morphologically rich languages. Many downstream language

[1] *Stem* and *lemma* do not necessarily refer to the same output due to applying different algorithms (i.e., stemmer and lemmatizer), but for this example, they converge since the verb root does not change when combined with other morphemes. However, if the word form is *flies*, the lemma is *fly* and the stem is *fli*. See *Porter stemmer* (Porter, 1980). Note that corresponding linguistic terms might not cove these cases.

processing tasks in an NLP pipeline such as syntactic and semantic parsing would require knowledge on certain linguistic properties encoded in words, and this can only be revealed by doing some kind of morphological analysis first. Therefore, computational models of morphology constitute the foundation of many low and high-level NLP applications and systems such as *part-of-speech tagging* (POS), *lemmatization, text-to-speech, spelling correction, machine translation, dialog systems, hyphenation*, and *stemming* for *information retrieval*.

In the literature, there are several examples proving that significant improvements are achieved for different applications in NLP (e.g., machine translation and speech recognition) when being trained on morphemes or other smaller string units than words[2] such as sub-words. It is claimed that such methods decrease the effects of morphological variations (Bojanowski, Grave, Joulin, & Mikolov 2017; Kirchhoff, Vergyri, Bilmes, Duh, & Stolcke, 2006; Sak, Saraçlar, & Güngör, 2010; Sennrich, Haddow, & Birch, 2016).

In the case of statistical or data-driven machine translation, a translation model without a module doing morphological analysis will raise *data sparsity* problems due to limited training data and also lead to untranslated *out-of-vocabulary* (OOV) words in output translations due to high diversity observed in languages (Singla, Sachdeva, Bangalore, Sharma, & Yadav, 2014). These problems ultimately undermine the performance of the system. However, more general representations focusing on word stems (e.g., *cat* for *cats*) will decrease the data sparsity while providing better statistics on training data (Koehn & Hoang, 2007), and decrease the number of OOV or unknown words in automatic translations[3].

The rich inflectional morphology of highly agglutinative languages like Finnish, Turkish, and Laz[4] produces not only a significant number of word forms but also very complex ones such as the one given in 2. To be able to process such a

[2] Word tokenizers usually determine word boundaries by looking at white space and punctuation but with certain exceptions based on language (e.g., Chinese).

[3] Morphological generation as another task in computational morphology is also a part of a machine translation system that produces word forms in the target language with provided syntactic and semantic properties.

[4] Laz exhibits both agglutinative and templatic (non-concatenative) patterns.

word or to make ‘sense’ of it, a computer requires access to a lexicon that is either big enough to cover all possible word forms in the language (i.e., a word-form lexicon) or created with morphemes and rules to combine them (i.e., morpheme-based lexicon). Generating and listing all the word forms manually would be very inefficient and time-consuming if not beyond one’s capacity to do so, especially when there is not enough data to train a machine learner to be able to process such word forms given in 2. Sproat and Stethem (1992) argues for the practical utility of computational morphology and state that simply expanding the dictionary or lexicon by manual addition of word forms is ‘wrong’ from a scientific point of view and fails to make use of ‘regularities’.

(2) *m-i-t’ax-ap-ur-t’-a-s-ert’u*
D-P.1-VAL:APPL-BREAK-CAUS-TS-IMPF-SUBJ-PRS.3SG-AUX
I would have broken it (by then).

Source: Öztürk and Pöchtrager (2011)

2.3 Methods and approaches to computational morphology

As for any other problem/task in NLP or CL, there are mainly two different approaches for developing an automatic morphological analyzer, namely rule-based and statistical/data-driven.

While rule-based methods require hand-crafted rules by the developer (e.g., two-level morphology (Koskenniemi, 1983)), statistical methods, as the name suggests, rely on statistics on some amount of training data to be able to access morphological rules/paradigms. Statistical methods can be further divided into supervised and unsupervised, depending on whether or not the training data fed into the machine learner is annotated or not.

Among others, unsupervised learning of morphology can be favored due to being adaptable to changes across languages and not requiring hand-crafted rules or manually annotated data which can necessitate a significant amount of knowledge, time, and money. However, one of the drawbacks while working with low-resource languages is clearly the small amount of data to begin with (written or spoken, annotated or non-annotated) (Hammam, 2008). Additionally, spoken languages or languages with several dialects also create problems due to the lack of a standard orthography. 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 rule-based morphosyntactic modeling for annotating small language data. In 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 and future development for computer-assisted language learning systems.

Rule-based approaches generally require three main components as a lexicon consisting of all the morphemes in the language including word stems and affixes, morphotactics to define how to combine these morphemes, and morphophonemics to define morphophonological changes when combining these morphemes.

2.3.1 Finite-state networks

Finite-state networks (FSN) are mathematically well-defined and have been used in computer science to define regular languages and relations (Hopcroft & Ullman, 1969) and they have been successfully applied to many language processing tasks beyond morphological analysis or generation to tokenization or shallow syntactic parsing (see Karttunen, Chanod, Grefenstette, and Schiller (1997)).

FSNs can be defined as directed graphs composed of *states* (one starting state and any number of accepting states) and labeled *arcs* (see Figures 1 and 2). If arcs are labeled on only one side, this makes the network a simple *automaton* as seen in Figure 1. If there are symbol pairs labeled on arcs, then the network is called a *transducer*. Therefore, two types of finite-state networks can be defined as finite-state automata (FSA) and finite-state transducers (FST). The former defines a regular language and determines if a string is accepted or not. The latter defines relations between strings in two different languages that allow mapping from one language to the other working as *bidirectional*. Two different languages here in fact refer to the lexical and the surface representation of a word.

The graph in Figure 1 simply shows the plural inflection on Laz nouns and is defined as a simple FSA in which the double-circled states, namely 2 and 3 are accepting states. This means that a noun stem on its own is accepted as well as when inflected with a plural suffix.

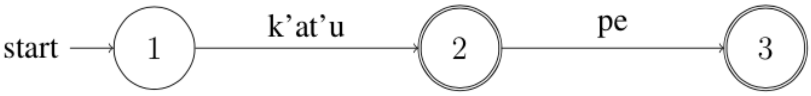


Figure 1. Finite-state automaton for Laz plural inflection (simplified)

The graph in Figure 2 extends the FSA in Figure 1 by adding symbol pairs to introduce the change of *i* to *e* in noun stems such as *limci* 'evening' when followed by plural *-pe*. Note that states 5 and 6 are not accepting states. This means that only after *i* is realized as *i* and forms a noun with *limc* or realized as *e and* followed by *-pe*, the word will be accepted as the *path* ends at the final accepting state 3. State 4 and 7 are dummy states just to mark the word with the POS tag <N>.

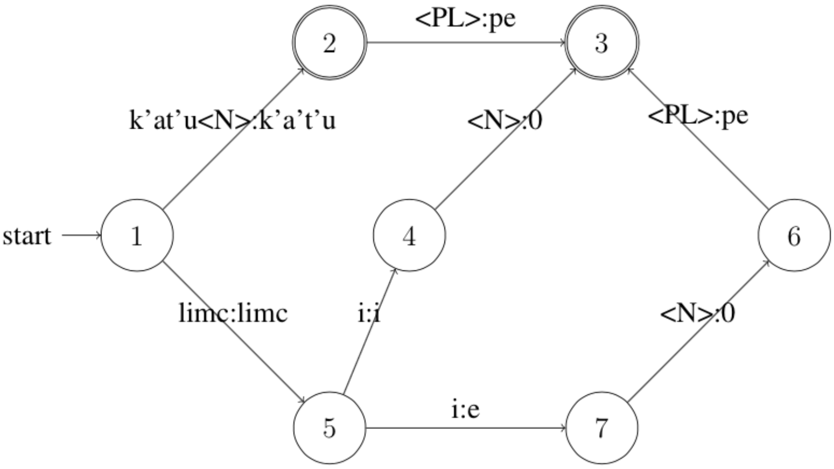


Figure 2. Finite-state transducer for Laz plural inflection showing *i* and *e* variation for noun stems ending in *i* (simplified)

As Figure 1 illustrates, a finite-state automaton can be used to model morphotactics only and function as an *acceptor* (Sproat & Stethem, 1992). This means that it can check if a given string (e.g., a word form) exists in this particular regular language by doing symbol matching between the input string and the string on the arcs. If the path ends in an accepting state, the string is accepted and found to exist in the language. Based on this, the language is formalized as an FSA in Figure 1 only has *k'at'u* and *k'a't'upe* in its lexicon. Such networks can be used in spell-checkers if they include an extensive amount of *lemmas*[5] (word stems) and also the morphotactics are well-defined.

Finite-state transducers have more functionality than a simple FSA in Figure 1. Since they are bidirectional due to having pairs of symbols on the arcs that allow mapping between lexical and surface forms as seen in Figure 2, they can work both as a recognizer and analyzer. Compare the possible output paths of the FSA in Figure 1 and the FST in Figure 2 given in 3, 4 and 5.

(3) FSA containing one symbol
k'at'upe

(4) FST containing symbol pairs
Lexical: k'at'u<N><PL>
Surface: k'at'upe

(5) FST containing symbol pairs
Lexical: limci<N><PL>
Surface: limcepe

2.3.2 Two-level morphology

Two-level morphology formalized by Koskenniemi (1983) with two-level referring

[5] The term *lemma* is the canonical form for a set of word forms that share the same core meaning but are inflected with different morphosyntactic information. It is also considered as the 'citation form' of an abstract notion *lexeme* (Blevins, 2016)

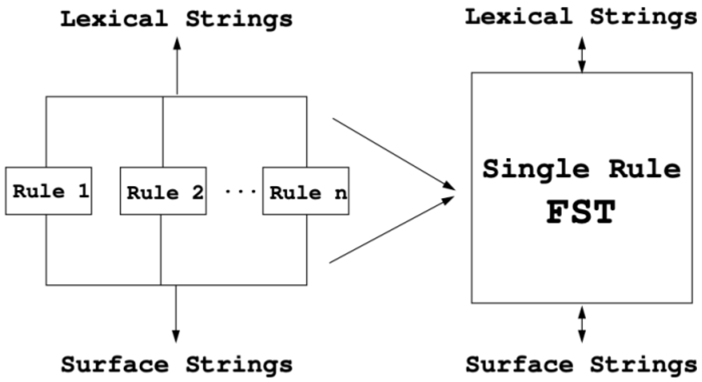


Figure 3. Two-level rules compiled into one single FST
Source: Karttunen and Beesley (2001), p. 11

to the lexical and surface representation of a word defines two-level phonological rules being applied in parallel to get the surface form directly without an intermediary stage resulting from sequential rewrite rules (Karttunen & Beesley, 2001). Based on this model, each two-level rule is defined as a finite-state transducer, and all two-level rules (or transducers) are applied at the same time in parallel, resulting in a single FST that takes a lexical form and outputs a surface form or vice versa.

Two-level morphology introduces the idea of *archiphoneme* (similar to the notion of a *phoneme* in linguistics) used on the lexical side. The lexical form referred to in two-level morphology is given in 7. With two-level rules such as the one in 6, this lexical form maps to its surface form seen in 7.

(6) Two-level rule for archiphoneme ‘I’
I -> e _ / pe

(7) FST compiled with two-level rules
Lexical: limcIpe
Surface: limcepe

At this stage, the idea of constructing lexicons using FSTs was not present (Karttunen & Beesley, 2001). Two-level morphology is only interested in writing phonological rules. With lexicon FSTs seen in Figure 4, it became possible to create all possible lexical forms in the language. Later, this lexicon is ‘composed’ with

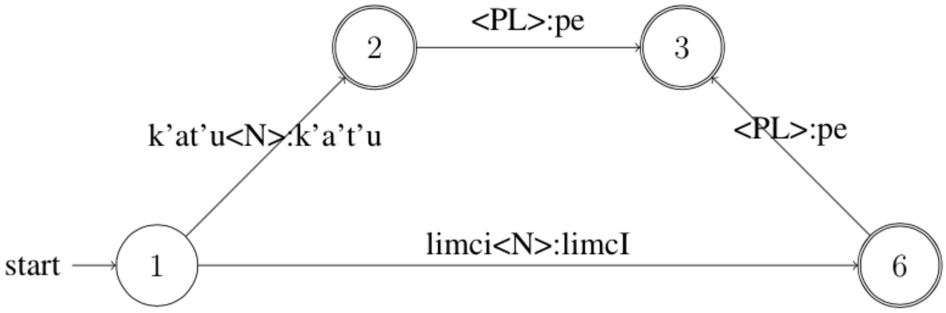


Figure 4. Lexicon finite-state transducer for Laz plural inflection introducing archiphoneme ‘I’ (simplified)

two-level rules into one single transducer which ultimately analyzes a word form into its morphosyntactic form or generates a word form from its morphosyntactic form (bidirectionality). One possible output path of the *lexicon* FST described in Figure 4 is given in 8. While one side of symbol pairs produces morphosyntactic forms (usually defined as <lemma, POS tag, morphosyntactic tags>, the other side produces lexical forms.

(8) Lexicon FST

Morpho-syntactic: limci<N><PL>

Lexical: limcIpe

Before several implementations came out to compile two-level rules into transducers, finite-state morphologies were compiled by hand (Karttunen & Beesley, 2001). Currently, there are several implementations such as Xerox (Beesley & Karttunen, 2003), Foma (Hulden, 2009), openFST (Allauzen, Riley, Schalkwyk, Skut, & Mohri, 2007), and HFST (Linden, Silfverberg, Axelson, Hardwick, & Pirinen, 2011). These implementations provide tools to compile both morphotactics and morphophonemic rules written with regular expressions in specific formalism such as *lexc* (lexicon compiler) and *twolc* (two-level compiler) into transducers.

2.4 Conclusion

In this chapter, I introduced why computational morphology matters and in what ways doing morphological analysis helps certain tasks and applications in NLP and CL. With morphologically rich languages, it is highly possible to observe word form variations in natural language data. There are several methods to deal with these variations but deciding on which one is the best depends on different conditions such as the amount and quality of language data. Therefore, although unsupervised or supervised machine learning methods are usually preferred since they require less time and effort, when there is no or little data, a ruled-based system could potentially serve better compared to state-of-the-art statistical methods in the field.

A ruled-based morphological analyzer requires three components, namely a lexicon, morphotactic rules, and morphophonological rules. The analyzer for this study is developed with finite-state networks and Koskenniemi (1983)'s two-level morphology.

CHAPTER 3

PAZAR LAZ

3.1 Introduction

This chapter will provide some background information on Laz and its current status 3.2, and dialects of Laz in 3.2.3. Later, it will introduce parts of Laz grammar that are fundamental to this study and constitute the main components of the morphological analyzer, starting with orthography in 3.3, phonetics and phonology in accordance with orthography in 3.4, morphotactics and morphosyntax in 3.5, and finally morphophonology in 3.6.

Because of the nature of this study, the main focus will be on descriptive grammar rather than theoretical work on Pazar Laz although I still refer to theoretical studies to clarify certain aspects of Pazar Laz grammar throughout this section.

3.2 Background

Laz language or *Lazuri*[1], which is mainly spoken on the northeastern coastline of Turkey and also on the Turkey borderline of Georgia has been recorded as a 'definitely endangered[2]' language in UNESCO Atlas of the World's Languages in Danger. It belongs to the South Caucasian language family (also called Kartvelian[3] in some literature) along with Svan, Mingrelian, and Georgian.

3.2.1 Information on Laz speakers

The number of Laz speakers were last officially reported as 85.108 including both native speakers and second language learners of Laz after the 1965 Turkey census.

[1] *Lazuri* is the term Laz people use.

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

[3] The term *Kartvelian* originally comes from *k'art'velebi* 'Georgians' in the Georgian language and it came into use based on the assumption of Georgian being the 'dominant' member of the language family (Boeder, 2005) and Georgian does not have an umbrella term to include all four languages in the language family (Hewitt, 2006).

Feurstein (1983) estimated this number as 250.000 for all Laz speakers in the world as in 1983 while Moseley (2010) provides a number ranging between 130,000 and 150,000 according to accounts reported in UNESCO Atlas of the World's Languages in Danger as in 2001[4].

Laz people are also Laz-Turkish bilinguals. Especially younger generations use Turkish more than Laz because of several social, economic, and political factors developed over the years the main one among which is Turkish being the medium of instruction in the country and the dominant language used in general public settings like workplaces[5] (see Kutscher (2008) for more details). Therefore, the use of Laz has been mostly restricted to the households or circle of friends and families in the region.

3.2.2 Development of a writing system

Until the 1920s Laz was a spoken language with only some written collection of Laz grammar and folklore studies. In 1928, İskender Ts'itaşi became the pioneer in developing a writing system for Laz based on the Latin alphabet in the Union of Soviet Socialist Republic, and later, Fahri Lazoğlu and Wolfgang Feurstein introduced and released 'Lazuri Alboni' (Laz Alphabet) to the public in Germany in 1984[6] (Boeder, 2005). Only after the 1990s, Laz people living in Turkey started using the alphabet and publishing literary works in Laz[7] when Laz intellectuals released Laz alphabet once more in a magazine called *Ogni* in Turkey in 1993 and 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 (Haznedar, 2018; Kavaklı, 2015).

[4] There are no up-to-date official sources that could give precise information on the current number of Laz speakers residing in Turkey or around the world due to lack of census on Laz speaking people or Laz ethnic groups in general; therefore, there are only estimations of this number found in the literature that range from as low as 20.000 to 500.000 (Kutscher, 2008; Lacroix, 2018; Salminen, 2007; Sarı, 2017).

[5] Some Laz people even accept Turkish as their native language rather than Laz (Haznedar, 2018).

[6] Including alphabets used in recent studies, different writing systems will be compared in section 3.3 , which discusses the orthography in detail.

[7] Political uneasiness and the rise of Turkish nationalism in Turkey during the period between 1930 and 1990 led to governmental policies enforcing the use of the Turkish language in public which adversely affected the use of several ethnic minority languages including Laz.

There was not much research on the lexicon and syntax of Laz until the end of 20th century when the first academic level studies started to emerge. Additionally, in 1999, the first dictionary for Laz (Turkish-Laz) was prepared and published by İsmail Bucak'lişi and Hasan Uzunhasanoğlu. In the following years, Bucak'lişi also published the first Laz grammar book (Kavaklı, 2015) and has been 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). Some world classics, such as The Little Prince, Romeo and Juliet, Snow White, Don Quixote, and Pinocchio were also translated into Laz. There have also been some short-lived journals and newspapers which aimed to publicize Laz language and culture.

3.2.3 Laz dialects

There are eight dialects of Laz including subdialects, none of which is considered to be the normative or 'standard'(Kavaklı, 2015; Kutscher, 2008). Bucak'lişi and Kojima (2003) specifies these dialects as Pazar (*Atina*), Çamlıhemşin (*Furthunaşi gamayona*), Fındıklı (*Viwe*), Arhavi (*Arkabi*), Ardeşen (*Arthaşeni*), Hopa (*Xopa*), Borçka-İçkale (*Çxala*) and Sapanca[8] based on the varieties of Laz spoken in different villages and divides them into two main groups as the Western dialects (*Gyulva*) consisting of Pazar, Çamlıhemşin, Ardeşen and the Eastern dialects (*Yulva*) consisting of Fındıklı, Arhavi, Hopa, and Borçka-İçkale [9].

The categorization above will be maintained for the lexicon because the dictionary used in this study, Bucak'lişi, Uzunhasanoğlu, and Aleksiva (2007) is prepared based on this division. It should be noted that some of these dialects are not regarded as separate but only subdialects under larger dialect groups. Lacroix (2018) states that the 'mutual intelligibility' for subdialects is not an issue although it might

[8] The Laz names of the dialects are given inside the parentheses.

[9] This division excludes 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.

not be the case among dialect groups. While Lacroix (2018) divides them into three main dialects as Hopa (including Borçka-İçkale), Fındıklı-Arhavi and Pazar (including Ardeşen) which are also accepted by Asillazanci (2018) and Ç'ikobava (1936) and Marr (1910). Kutscher (2001) and Öztürk (2019b) use a four-way distinction separating Ardeşen and Pazar dialect due to differences in case marking. It is also possible to encounter studies that assume five major dialects of Laz. Even though underlyingly the structure of these dialects is the same, they show lexical and morphosyntactic, as well as phonological differences[10].

For this thesis, I chose to build (or start building) the morphological analyzer on the Pazar Dialect since there is now considerable amount of research done on morphosyntactic aspects of Pazar Laz.

3.3 Orthography

Orthographically, I have adopted the 1984 Lazoğlu alphabet given in Tables 1 and 2 which is an extended version of the Turkish alphabet based on Latin letters, specifically for the development of the morphological analyzer although here, I follow the Lacroix (2009) and Öztürk and Pöchtrager (2011)'s version since it is more intuitive to me, especially considering the resemblance to the IPA's sound symbols. The reason why I chose the Lazoğlu alphabet is the fact that it is commonly used by Laz speakers and therefore, most written texts follow that convention.

See Bucak'lişi and Kojima (2003) for details on different alphabets used for Laz and also Asillazanci (2018) for original documents of Ts'itaşi (1932) and Lazoğlu and Feurstein (1984) alphabets.

3.4 Phonetics and phonology

I have already given the partial sound system of Pazar Laz in Tables 1 and 2 that show the sound correspondences to the letters used in various writing systems developed over the years during the last century. See Öztürk and Pöchtrager (2011) pages 5-11

[10] See Bucak'lişi and Kojima (2003), a descriptive grammar book for an extensive analysis on all eight dialects of Laz defined above and specific differences among them

Table 1. The 22 letters of the Laz Alphabet

Ts’itaşi (1932)	Lazoğlu and Feurstein (1984)	Bucak’lişi et al. (2007)	Lacroix (2009) Öztürk and Pöchtrager (2011)	IPA Symbol
a	a	a	a	[a]
b	b	b	b	[b]
ç	ç	ç	ç	[tʃ^{h}]
d	d	d	d	[d]
e	e	e	e	[e]
f	f	f	f	[f]
g	g	g	g	[g]
h	h	h	h	[h]
i	i	i	i	[i]
l	l	l	l	[l]
m	m	m	m	[m]
n	n	n	n	[n]
o	o	o	o	[o]
p	p	p	p	[p^{h}]
r	r	r	r	[r]
s	s	s	s	[s]
ş	ş	ş	ş	[ʃ]
t	t	t	t	[t^{h}]
u	u	u	u	[u]
x	x	x	x	[x]
v	v	v	v	[v]
y	y	y	y	[j]
z	z	z	z	[z]

for a full overview of the sounds found in Pazar Laz which include five additional labio-velar sounds as [k^{h}w], [gw], [k$^{’}$w], [xw] and [ɣw].

In Pazar Laz, Öztürk and Pöchtrager (2011) defines a five-vowel system composing of [a], [e], [u], [o] and [i][11] which is found in other Laz dialects as well

Table 2. The 12 Letters of the Laz Alphabet

Ts’itaşi (1932)	Lazoğlu and Feurstein (1984)	Bucak’lişi et al. (2007)	Lacroix (2009) Öztürk and Pöchtrager (2011)	IPA Symbol
c	ʒ	ts	ts	[ts]
ch	ʒ̆	tz	ts’	[ts’]
çh	č̹	çh	ç’	[tʃ’]
ıo	ğ	ğ	ğ	[ɣ]
j	c	c	c	[dʒ]
k	ǩ	q	k’	[k’]
ph	p̆	ph	p’	[p’]
q	k	k	k	[kʰ]
th	ť	th	t’	[t’]
z̭	j	j	j	[ʒ]
ʒ	ž	zh	dz	[dz]
-	q	q’	q’	[q’]

Note: Although the uvular ejective sound [q’] is lost in Pazar Laz, Öztürk and Pöchtrager (2011) accepts the existence of the sound in Hopa dialect of Laz; therefore, I added the letter in the Öztürk and Pöchtrager (2011) alphabet.

as in other South Caucasian Languages (Beguš, 2019). Additionally, there are 34 sounds in the consonantal inventory of Pazar Laz, including labialized velar sounds[12].

Beguš (2019) notes ‘while phonemic inventories are similar across Caucasian languages, the phonetic realization of phonologically identical segments can differ substantially, even in closely related languages or even between different dialects of the same language’, referring to ejective[13] sounds which I will discuss in the next

[11] Beguš (2019) uses the following phonemic inventory for Laz respectively: /a, e, u, o and i/.

[12] These sounds are considered as one single unit in Öztürk and Pöchtrager (2011) as opposed to the consonantal cluster analysis in (Bucak’lişi & Kojima, 2003; Lacroix, 2009). According to Catford (1977), they are typically found in phonemic inventories of Northwest Caucasian languages but ‘sporadically’ outside these languages.

[13] *Ejectives* are voiceless stops or obstruents produced with the complete closure of the glottis released with an upward movement of the glottis in addition to a vocal tract constriction which mainly characterizes pulmonic or plain stops. They are usually perceived as without aspiration and with a creaky voice on the following vowel but there is not one specific acoustic parameter that distinguishes ejectives from other stops.

section. In addition, as mentioned before, Laz has no standard form, written or spoken (Kutscher, 2008). This makes it highly possible to encounter orthographic variations that mark acoustic differences for the same word in the language, not just between dialects but sometimes between people speaking the same dialect.

(9) Laz words for ‘anchovy’, also showing the variation between velar ejective [k’] and voiceless-aspirated stop [k^{h}][14].

kapça	[k^{h}aptʃʰa]	(Fındıklı- Arhavi)
k’apça	[k’aptʃʰa]	(West dialects)
k’apşa	[k’apʃa]	(Ardeşen)
kapçia	[k^{h}aptʃʰia]	(Borçka- İçkale)
kapşia / kapsia / kapşira	[k^{h}apʃia] / [k^{h}apsia]/ k^{h}apʃira]	(Hopa)

Source: Bucak’lişi et al. (2007), p. 468

In 9, seven different representations for the word ‘anchovy’ in Laz, three of which are from the Hopa dialect only can be seen. Although I provided the phonetic representations inside square brackets based on the orthography and correspondences given in Tables 1 and 2 between Laz letters and IPA symbols, it should be noted that the transcription may or may not reflect the actual or exact phonetic realization of words in actual speech. Since what matters for the analyzer is the orthographical representations and changes for word forms rather than abstract phonetic or phonemic representations of sounds, I will here touch upon where phonetics and phonology affect orthography, argue that the lexicon should have words not only from Pazar Laz but from all dialects as a way to best deal with all these variations when developing the analyzer. In this respect, I will discuss only one case of variation as an example, namely free variation with stops with different laryngeal properties intra- and inter-dialectically.

Laz differentiates stops in terms of their laryngeal properties, namely voiced, voiceless-aspirated, and ejective stops (Anderson, 1963; Beguš, 2019; Lacroix, 2009;

[14] Example 10 and 12 show variations between other sounds as well, which I will not discuss here because of the scope of this thesis work.

Öztürk & Pöchtrager, 2011). Some dialects display ejection at three places of articulation as bilabial, alveolar, and velar, and some like Hopa and Borçka-İçkale at four places of articulation with the addition of uvular[15], also seen inTable 3. Ejectives are in fact very typical in Caucasian languages (Catford, 1977) although Grawunder, Simpson, and Khalilov (2010) notes that each Caucasian language uses this feature to different degrees, strong or weak[16].

Table 3. The Inventory of Laz Stops

	Bilabial	Alveolar	Velar	Uvular (Hopa and Borçka-İçkale)
Aspirated	*p^h*	*t^h*	*k^h*	
Ejective	*p'*	*t'*	*k'*	*q'*
Voiced	*b*	*d*	*g*	

Stops with different laryngeal properties are the sounds that seem to also mark dialectical differences in the dictionary (Bucak'lişi et al., 2007). Although every dialect of Laz uses all three types of stops, at least in writing, where to use one as opposed to others changes from dialect to dialect, and even speaker to speaker (also from my observations with speakers from both same and different dialects). These changes can be seen in the examples below in 9, 10, 11, and 12. The dialects that the variations belong to are given inside the parenthesis and the phonetic/acoustic representations inside square brackets.

(9) Laz words for 'bad', showing the variation between bilabial ejective [p'] and bilabial voiced stop [b] in the word-initial position

p'at'i	[p'at'i]	(Arhavi)
p'iat'i / p'eat'i	[p'iat'i]/[p'eat'i][	(Ardeşen)
biat'i	[biat'i]	(Hopa)

Source: Bucak'lişi et al. (2007), p. 1048

[15] Bucak'lişi and Kojima (2003) also defines three types of ejectives at uvular as an ejective fricative [X'] and also as an affricate [qX'] together with the uvular ejective stop [q'].

[16] This has led to different measurement criteria for ejectives in the literature, and the growing of a research area for the acoustic *typology* of ejectives across Caucasian languages (Kingston, 1985).

(10) Laz words for ‘honeybee’, and the variation between bilabial ejective [p’], voiceless-aspirated [p] and voiced stop [b] in the word-initial position

p’ut’uci	[p’ut’udʒi]	(Atina)
but’uci	[but’udʒi]	(Çamlıhemşin)
but’uji	[but’uʒi]	(Ardeşen)
but’k’uci	[but’k’udʒi]	(Fındıklı)
put’ut’k’i	[p^{h}ut’ut’k’i]	(Megrelian)

Source: Bucak’lişi et al. (2007), p. 58

(11) Laz words for ‘pull’, and the variation between alveolar and velar ejectives [t’] and [k’] and [t] and [k] respectively in the intervocalic position

ost’ik’u	[ost’ik’u]	(West dialects)
ostiku	[ost^{h}ik^{h}u]	(East dialects)

Source: Bucak’lişi et al. (2007), p. 967

Fallon (2013) argues that the observed variations in Laz might be a case of complete deglottalization[17] in some dialects or ‘deglottalization in free variation’. Further, he also suggests that the variations could be simple ‘doublets’ (or triplets in some cases), referring to Turkish borrowings in Laz like in 13 below, arguing that ‘the nonglottalized case may be adopted without nativization’ because of Laz speakers being bilinguals with the Turkish language.

The data found in the dictionary and also in corpora here confirms that not only Turkish borrowings show this variation, but also native Laz words clearly display this ‘free variation’ between Laz stops with different laryngeal features. As the reason behind this issue, Haig and Khan (2018) discusses that the geographic isolation from languages (e.g., Georgian) that contain ejective stops leads to ‘weakening or reduction’ of the glottalic feature in the language as well as the influence of Turkish in which there are no ejective sounds. If the ejective occurs in rapid speech, this also affects the glottal feature to be less sound and as Fallon (2013) describes it, ‘ejectives seem to have merged with a voiceless series’ . In fact,

[17] *Deglottalization* means that a glottal sound like ejective loses its secondary constriction at the glottis while preserving the primary vocal track constriction and becomes a plain or pulmonic stop.

Lacroix (2018) points out that Laz ejectives do not have strong ejection as those in Georgian. Öztürk and Pöchtrager (2011) reports that ejectives in Laz are ‘phonetically only mildly ejective in character’.

(12) Laz words for ‘fate’[18], the variation between velar ejective [k’] and velar voiced [g], voiceless-aspirated stop [k^h and also velar voiceless fricative [x] observed in Megrelian

igbali	[igbali]	(Hopa)
ikbali	[ik^hbali]	(Fındıklı)
ik’bali	[ik’bali]	(Pazar)
iğbali	[ixbali]	(Megrelian)

Source: Bucak’lişi et al. (2007), p. 449

Although the dictionary provides an extensive amount of information on these kinds of orthographical variations, specific for each dialect, in a way creating ‘standards’, I decided not to limit the lexicon to the words defined only for Pazar Laz in the dictionary since all these issues discussed are very present in Laz and transcriptions reflect the phonetic realization of sounds in speech (Lacroix, 2009).

I received transcriptions from a native speaker of Laz from the Hopa dialect. His transcription involves words categorized under other dialects such as *p’at’i* [p’at’i] ‘bad/hurting’ (Arhavi) instead of *biati* [biati] (Hopa) as well as words not defined in the dictionary such as *kapşiya* instead of any word given in Example 9. He sometimes uses voiceless-aspirated and ejective stops in free variation for some words when used in sentences such as [k] and [k’] in *stik-eri* and *stk’-eri*[19] ‘pulled (grass)’ (participle form), and [p] and [p’] in *i-piç-up-s* and *i-p’iç-up-s*[20] (3rd singular) ‘sleeps/sleeping’, respectively but uses only one form for those words as ‘o-stik’-u’[21] and ‘o-p’iç-u’ when in isolation and in their infinitival forms.

These examples prove how transcriptions may not follow ‘standards’ in the case of an understudied and endangered language with no standard written form like

[18] This word comes from the Arabic word *ikbal* and it is also commonly used in Turkish.

[19] *-eri* suffix is the participle suffix.

[20] *i-* is the valency-related or version vowel, *-up* is thematic suffix marking progressive aspect, and *-s* marking 3rd person singular.

[21] In dictionary, only *ost’ik’u* and *ostiku* forms are provided although the speaker uses *ostik’u*.

Laz because phonetic characteristics of these sounds in running speech and speakers' perception of them affect their transcriptions [22]. One might think that in this case categorizing dialects under larger groups like west and east dialects would be a better idea, but it is not unexpected to see some borrowings between west and east dialects as well. In fact, the corpus I used for this study for Pazar Laz contains words from eastern dialects like *açkveneri*[23] (Hopa) 'next time'. Therefore I decided to still keep the dialectical categorization as provided in the dictionary but making use of the *most* general information and let the whole lexicon open for the analyzer developed on Pazar Laz morphotactics and morphophonological rules.

3.5 Morphotactics and morphosyntax

In this section, I will introduce how morphemes, the smallest meaningful units are combined to form morphologically complex word forms in Pazar Laz as well as certain morphosyntactic properties of Pazar Laz. There are suffixes and prefixes in Laz which are sometimes used together forming long-distance dependencies on the word level. Additionally, Laz exhibits non-concatenative, templatic morphology similar to languages Arabic and Hebrew, specifically for inflecting verb class (Atlamaz, 2013).

This part of Pazar Laz grammar also constitutes the backbone of the morphological analyzer with root forms or lexemes defined in the lexicon.

3.5.1 Inflectional Morphology

Inflectional morphology is interested in grammatical relations between words in a sentence or in other words, morphosyntactic properties of word forms. Inflectional affixes for nouns and verbs are discussed in this section.

[22] These changes in transcriptions could be also simple spelling mistakes. However, because of the nature of ejectives and how they are phonologically contrastive with other stop series, yet still used somewhat 'interchangeably' with them in a more sporadic rather than systematic way, it is impossible to judge if they are correct or not. Additionally, I am in no place to suggest a 'standard' form for these words in any case since I consider myself as an observer but neither a (native) speaker of Laz or an etymologist trained in these studies.

[23] This word is a compound composing of *a* 'one' and *çkva* (East dialects) / *şk'va* (Western dialects) 'other'.

I will define certain morphotactic rules in this section, mostly regarding the verbal complex which deviate from simple concatenative morphology since these morphotactics require special treatment in FSTs (i.e., Flag diacritics).

3.5.1.1 Nouns

Compared to verb conjugation which will be discussed in Section 3.5.1.2, nominal inflection is rather simple in terms of morphotactics but not so much morphosyntactically and semantically. Pazar Laz marks nouns with case and number but not gender in general. Case markers and plural marker *-pe* are shown in Table 4 inflecting the lexeme *bere* 'child'.

Table 4. Case and Number Inflection for *bere* 'child' in Pazar Laz

	Singular (SG)	Plural (PL)
Nominative (NOM)	bere-Ø	bere-pe-Ø
Ergative (ERG)	bere-k	bere-pe-k
Dative (DAT)	bere-s	bere-pe-s
Ablative (ABL)	bere-şe	bere-pe-şe
Allative (ALL)	bere-şe	bere-pe-şe
Genitive (GEN)	bere-şi	bere-pe-şe
Instrumental (INST)	bere-te	bere-pe-te

Plural nouns are expressed by adding the suffix *-pe* in Pazar Laz[24]. Additionally, some lexemes ending in *a* take *-lepe* but not all. See examples in 14.

(14)				
	dida	*'old woman'*	*didalepe*	*'old women'*
	toli	'eye'	tolepe	'eyes'
	nca	'tree'	ncalepe	'trees'

Morphotactic constraint 1: Only certain lexemes ending with *a* can be marked with the *-lepe* plural marking[25].

[24] If the root ends with *i*, *i* becomes *e* before *-pe*. This rule is added as a morphophonological rule.

[25] The dictionary by Bucak'lişi et al. (2007) provides this information in the word definitions. They are extracted and labeled based on the existence of the *-lepe* suffix in the word entries.

Table 5. Possessive Pronouns in Pazar Laz

	Singular (SG)	Plural (PL)
1st person	şk'imi	şk'uni
2nd person	sk'ani	t'k'ani
3rd person	himuşi	hinuşi

There are also some noun phrases composed with possessive pronouns and postpositions in which the noun stem and possessive pronouns or postpositions can be found to be adjacent to one another seen in 15. Therefore, the morphotactics for nouns include possessive pronouns and postpositions as well. The concatenation follows this ordering in 16.

(15) *Nana şk'imi/ Nana-şk'imi /Nanaşk'imi*
house my
'My house'

(16) noun stem + plurality + possessive + postposition + case + additive

Possessive pronouns are given in Table 5. In addition to their adjacent nature, they can also stand-alone pre- or post-noun stem. Pazar Laz additionally has *muşi* 'her/his' and *nişi* 'their' only occurring post-noun stem.

Similar to possessive pronouns, postpositions can occur on their own or combine with noun stems such as the example in 17. The example in 18 also shows a possible combination of the noun stem with both a possessive and a postposition.

(17) *mts'k'upişk'ala / mts'k'upi-şk'ala / mts'k'upi şk'ala*
dark-with
'with the dark'

Source: Bucak'lişi et al. (2007)

(18) *bozomota-lepe-muşi-şk'ala*
daughter-PL-his/her-with
'with his/her daughter' Source: Bucak'lişi et al. (2007)

Finally, nouns can be inflected with the additive (ADD) marker *-ti* following case markers shown as below in 19.

(19) *Ali-Ø-ti* *mo-xt'-a-s-ere*
girl-NOM-ADD PRV-come-SUBJ-PRS.3SG-FUT
'Ali is coming, too.'

Source: Öztürk and Pöchtrager (2011)

There are seven case markers defined for Pazar Laz shown in Table 4. Although the functions of these markers on the syntactic level are not of interest for the morphological analyzer, it is important to have an idea about how the case assignment works in Pazar Laz to understand the inflectional complexities inside the verbal complex, and also for more advanced syntactic or semantic parsing tasks following morphological analysis in any NLP pipeline.

3.5.1.1.1 Typology of Laz: Case assignment and split-ergativity

One important aspect of the case marking in Laz is that case is not determined based on grammatical functions of syntactic elements such as the subject or object (e.g., nominative-accusative alignment) or solely on the transitivity of the verb (e.g., absolutive-ergative alignment), but on the semantic roles (thematic roles) of these elements such as *agent, patient*, *theme*, *experiencer*, *benefactor*, *possessor* or *recipient*. This situation suggests the presence of *active* case alignment in Laz (Demirok, 2013).

As an additional note for the paper I co-authored for this project (Onal & Tyers, 2019), I would like to correct the mistake of categorizing Laz as an absolutive-ergative language in terms of morphosyntactic alignment since it also shows patterns from nominative-accusative languages[26]. In fact I would like to discuss here briefly that Laz exemplifies as a specific ergative language that exhibits a phenomenon called *split-ergativity* (Demirok, 2013; Dixon, 1987; Öztürk, 2013) while describing the nominative (NOM) and ergative (ERG) cases in Laz.

[26] Case assignment depends on the grammatical function of the noun phrase. If the noun phrase occupies the subject position, it takes the unmarked NOM case, and if the object position, then it takes the accusative (ACC) case.

Different from ergative-absolutive languages where the ERG case only surfaces on the agentive subject of transitive constructions while the subject in intransitive constructions is always unmarked with the absolutive or nominative case (unmarked), in the case of split-ergativity in Pazar Laz the agent or causer subject of one type of intransitives, namely unergatives is also marked with the ERG case in addition to the agent argument of transitives. Examples are given in 20 and 21.

(20) Intransitive (unaccusative, patient/theme subject)

K’oçi-Ø do-ğur-u
girl-NOM PRV-die-PST.3SG

‘The man died.’

(21) Intransitive (unergative, agent/causer subject)

K’oçi-k i-gzal-s
man-ERG VAL-walk-PRS.3SG

‘The man is walking.’

For more extensive discussion on ergativity in Laz, see Demirok (2013) and Gürpınar (2000).

Apart from the theme subject of unaccusative intransitive predicates, the NOM case also marks the patient object in certain transitive constructions while the ERG marks the subject seen in 22.

(22) Transitive

Bere-k kva-Ø do-k’an-am-s
child-ERG rock-NOM PRV-throw-TS-PRS.3SG

‘The child is throwing rock.’

3.5.1.1.2 Dative case

The use cases of the dative (DAT) marker are not as straightforward as the NOM and ERG cases (Demirok, 2013). There are a range of semantic roles it can attach to. These semantic roles could be lexically determined by the verb head or introduced with non-core arguments[27] as a result of *applicativization* or *causativization* phenomena which I will explain with examples later in this section, and also in Section 3.5.1.2 since the verb complex is inflected accordingly with applicative morphology when non-core arguments with the DAT case are introduced into the structure. Additionally, related to applicativization phenomena, there are *inversion* constructions as being defined in Öztürk and Pöchtrager (2011) with which *abilitative* and *deagentive* semantic roles are introduced with the DAT marking on the core arguments and applicative morphology on the verb complex (Demirok, 2013). These constructions add a 'circumstantial modality' without adding a non-core argument and changing the number of arguments in the structure. A similar process is also observed with present perfect constructions in Pazar Laz with the addition of 'experiential perfect' modality (Demirok, 2013; Öztürk & Pöchtrager, 2011; Öztürk, 2013).

First of all, I will show the inherent DAT marking on lexically selected *core* arguments. The following arguments bear the DAT case under this categorization followed by their example sentences;

- The subject bearing the *deagentive*[28] role of certain physiologicalintransitive verbs such as *-çind-* 'to sneeze' in 23

(23) *Bere-s a-çind-e-n.*
child-DAT VAL:APPL-sneeze-TS-PRS.3SG
'The child is (involuntarily) sneezing.'[29].
Source: Öztürk and Pöchtrager (2011)

[27] Non-core (applied) arguments are not part of the initial argument structure of the verb head but are introduced into the argument structure with certain operations. They usually assume the benefactive-malefactive semantic role, but they can also take recipient, goal, source, possessor and location roles.

[28] Deagentive role denotes involuntary action on part of the subject.

[29] This sentence with involuntary reading is considered to be an example of inversion according to Öztürk and Pöchtrager (2011) and Öztürk (2013) because of applicative morphology realizing as the pre-root version vowel (i.e., *a-*) with the DAT case on the subject and the thematic suffix *-e(r)* (supporting the suppressed, 'inversed' subject surfacing as the DAT argument) marked on the verb. However, since *bere* 'child' is seemingly the core argument of the verb *-çind-* 'to sneeze', I categorized this type of DAT marking under lexically selected DAT arguments. Interestingly, the volitional reading (non-applicative version) leads to a different structure where the subject is inflected with ERG and there is no applicative-related inflection, but the change of thematic suffix to *-um* (supporting the agentive subject) on the verb as seen in the sentence below in (i). Therefore, what is important for this study is that the analyzer should be able to cover both structures, specifically the different inflections on the verb *-çind-* 'to sneeze' rather than the DAT vs ERG inflection on the noun. In fact, the relation to the

- The subject bearing the *experiencer* role of certain transitive psychological verbs such as -*limb*- 'to love/like', -*şk'ur*- 'to fear', -*şin*- 'to remember', -*cer*- 'to believe', and -*ç'ondr*- 'to forget'[30]

(24) *Bere-s coğorepe-Ø a-limb-e-n.*
child-DAT dogs-NOM VAL:APPL-love-TS-PRS.3SG
'The child loves dogs.'

- The object bearing the *patient* role of certain transitive verbs such as -*şvel*- 'to help', -*xel*- 'to kiss', or -*nts'*- 'to touch'[31](Gürpınar, 2000), and the object lexically required to bear DAT as the 'inherent applied object' such as -*yox*- 'to call (at someone)', and -*yondr*- 'to wait (for somebody)' (Öztürk & Pöchtrager, 2011)

(25) *Arte-k bozomota-s gv-a-xel-u.*
Arte-ERG girl-DAT PRV-VAL-kiss-PST.3SG
'Arte kissed the girl.'

nominal inflection is crucial for more complex and advanced semantic or syntactic processing tasks such as *dependency parsing* (e.g., how to process both ERG and DAT arguments as the subject).

(i) *Bere-k çind-um-s.*
child-ERG sneeze-TS-PRS.3SG
'The child is (voluntarily) sneezing.'
Source: Öztürk and Pöchtrager (2011)

[30] Öztürk (2013) and Demirok (2013) argue that the subject of these verbs takes the *experiencer* role because it is the applied argument as a result of *applicativization* and applicative morphology marked over structures given in (ii) which shows the non-applicative version of the sentence in 24:

(ii) *Coğorepe-Ø i-limb-e-n.*
child-NOM VAL:PASS-love-TS-PRS.3SG
'Dogs are loved.'
Source: adapted from Demirok (2013)

Öztürk and Pöchtrager (2011) additionally discusses this issue under *inversion* (similar to structures with *deagentive* arguments given in 23) which is an operation suppressing the external argument (e.g., the subject) with the DAT, and marking the person value of the subject and the object in a 'reverse' order on the verb (e.g., marking the subject information preverbally due to this DAT marking). Therefore, the version vowels marked on these verbs are glossed as VAL:APPL, following the glossing in Öztürk and Pöchtrager (2011), referring to the applicative morphology. However, they are still considered 'inherent' because of the semantics and the core argument structure of the verb.

[31] Gürpınar (2000) explains the use of DAT on the *patient* instead of NOM for these verbs by referring to their low 'degree' of transitivity compared to verbs with NOM marked patients. DAT marked patients are not subject to change like NOM marked patients are.

(26) Inherent applied object

Nana-k bere-s u-yox-u.
mother-ERG child-DAT VAL:APPL-call-PST.3SG
‘The mother called at the child.’

Source: Öztürk (2021)

- The indirect object bearing the semantic role as the *recipient* of ditransitive verbs such as -*ç*- ‘to give’ n and -*ts’ir*- ‘show’.

(27) *Arte-k çitabi-Ø Fadime-s ko-me-ç-u.*
Arte-ERG book-NOM Fadime/Fatma-DAT PRV-PRV-give-PST.3SG
‘Arte gave the book to Fadime/Fatma.’

Unlike the high unpredictability of the DAT case on core arguments, the DAT case on non-core applied arguments is always predictable in Pazar Laz (Öztürk, 2013) although the semantic role of these arguments varies based on the verb stem and the discourse.

I will discuss non-core arguments under three different phenomena as (valency-increasing) applicativization, causativization, and inversion[32].

- Applicativization introduces a non-core object to the argument structure of the verb. The new argument can bear the semantic role of benefactor, recipient, location, or possessor and is always marked with the DAT case (Öztürk, 2013). It basically increases the number of arguments by one. 28, 29, and 30 are adapted from Demirok (2013).

(28) Non-applicative construction

Nana-k ar past’a-Ø ç’-u.
mother-ERG a cake-NOM bake-PST.3SG
‘The mother baked a cake.’

(29) Applicative construction: benefactor

[32] Inversion is categorized differently although in the literature it is mostly discussed in relation to applicativization because of shared applicative morphology and DAT applied arguments (i.e., present-perfect constructions (Öztürk & Pöchtrager, 2011))

Nana-k bere-s ar past'a-Ø u-ç'-u.
mother-ERG child-DAT a cake-NOM VAL:APPL-bake-PST.3SG
'The mother baked a cake for the child.'

(30) Applicative construction: possessor

Nana-s skiri-ø u-ğur-u.
mother-DAT child-NOM VAL:APPL-die-PST.3SG
'The mother's child died.'

The difference of an applicative construction from ditransitive predicates is the marking of the verb complex with pre-root version vowels (shown as VAL:APPL), specifically *i-/u-* or *a-*. While the verb complex in a ditransitive in 27 does not include a version vowel, the verb in an applicative construction given in 29 is marked with the version vowel *u-*. Their functions will be discussed in the section for verb conjugation in 3.5.1.2.

- Causativization is the operation of adding a *causee* argument to the structure. The DAT marking surfaces on the causee argument only when a transitive predicate is causativized but not an intransitive. The causativized version of 28 is given below in 31.

(31) *Nana-k bere-s ar past'a-Ø o-ç'v-ap-u.*
mother-ERG child-DAT a cake-NOM VAL:CAUS-bake-CAUS-PST.3SG
'The mother made the child bake a cake.'

- Inversion, unlike valency-increasing applicativization and causativization operations, does not add a new argument to the structure but causes the agentive subject to be suppressed/backgrounded and marked with the DAT instead of the ERG, and the ERG case, in fact, is never possible for these constructions (similar to valency-decreasing operations). Inversion is observed with certain modalities such as circumstantial modality introducing abilitative or deagentive roles, and experiential perfect[33] (Öztürk & Pöchtrager, 2011).

[33] Present perfect is also considered as a way to express modality, semantically adding the meaning that the subject has gained the experience in an event/action for an unspecified period of time, having started sometime in the past (Demirok, 2013; Öztürk & Pöchtrager, 2011). The way marking the subject with the DAT case and the corresponding applicative morphology on the verb also support this view. This issue will be discussed more in section 3.5.1.2. See the translations in English given in 33.

(32) Non-applicative construction: No inversion

Bere-k Lazuri-Ø d-i-gur-u.
child-ERG Laz-NOM PRV-VAL -learn-PST.3SG
'The child learnt Laz.'

(33) Applicative construction: Inversion (ambiguous betweenmodalities, ability and involuntary action)

Bere-s Lazuri-Ø dv-a-gur-u.
child-DAT Laz-NOM PRV-VAL:APPL-learn-TS-PST.3SG
'The child unintentionally learnt Laz.'/'The child was able to learn Laz.'

(34) Applicative construction: Inversion (experiential perfect construction)

Bere-s Lazuri-Ø d-u-gur-ap-u.
child-DAT Laz-NOM PRV-VAL:APPL-learn-CAUS-PST.3SG
'The child has learnt Laz.'

3.5.1.1.3 Other case markers

The ablative (ABL) case *-şe(n)* is used to mark the source. This could be the source of *fear* as in 35. The final [n] sound only surfaces when preceding the postposition *doni* 'since' which selects a noun phrase marked with the ABL (Öztürk & Pöchtrager, 2011).

(35) *Tanura-s mts'upi-şe a-şk'ur-in-e-n.*
Tabura-DAT dark-ABL VAL:APPL-fear-CAUS-TS-PRS.3SG
'Tanura is afraid of dark.'

Source: adapted from Bucak'lişi et al. (2007)

The allative (ALL) case *-şe* marks the direction or the goal as in 36:

(36) *Noğa-şe u-l-ur*
city.center-ALL VAL-go-TS
'He/she is going to the city center.'

Source: adapted from Bucak'lişi et al. (2007)

Lastly, the genitive marker -*şi* is used to mark the possessor in possessive constructions (e.g., *k'at'u-şi tolepe* 'cats' eyes') as well as the complements of certain spatial postpositions such as *oği* 'in front of' or *melenk'ale* 'opposite of' while the instrumental suffix -*te* is used to mark instruments as in 37.

(37)	İni	*ts'ari-te*	*dolonkuna-Ø*	*do-nax-u*
	cold	water-INST	laundry-NOM	PRV-wash-PST.3SG

'He/she washed the laundry with cold water.'

Source: adapted from Bucak'lişi et al. (2007)

3.5.1.2 Verb

As I discussed before in Section 3.5.1.1.1 regarding the *active* case alignment based on semantic roles, in terms of morphosyntactic alignment, Laz is an ergative language that also shows characteristics of nominative-accusative alignment (known as *split-ergativity*). In this section, I will talk about the consequences of such a system on the verbal complex or template which can be modified by previously discussed valency-related operations such as applicativization, causativization, and inversion with related applicative morphology. Concerning these, I will also discuss long-distance dependencies observed between verbal inflectional affixes and formalize them as morphotactic rules.

The verbal complex is compiled in Tables 6, 7, 8, and 9[34] in a concatenative manner. However, it should be noted that not all 16 positions presented in these tables surface at the same time, and only person marking is obligatory (Öztürk & Pöchtrager, 2011). Additionally, as Atlamaz (2013), Demirok (2013), and Öztürk (2021) argue, unlike the post-root affixes which are mostly agglutinative, the preverbal markers such as person prefixes or valency-related vowels 'compete for limited morphological slots', meaning that only one prefix per position is allowed although there might be more than one possible candidate for the position. This situation leads to the formation of a hierarchy between person markers to be able to rule the preverbal position. Demirok (2013) argues that 'syntactic hierarchy of

[34] The numbers mark the position in the verbal complex.

arguments’ based on their case markers (i.e., ERG, NOM, and DAT) and semantic roles is the reason behind this and defines this hierarchy as *object* > *subject* in the most general sense. Another competition is observed between valency-related vowels when there is more than one operation the predicate undergoes such as applicativization, causativization, and inversion, and the winner is determined according to a hierarchy between applicative constructions in the syntactic structure (see Öztürk (2013) for a more extensive discussion).

Table 6. Pre-Root Complex

Affirmative preverb -4	Spatial preverb -3	Person prefix -2	valency-related (version) vowel -1	Root 0
o-,ko-, do-,menda-	ama-, ce-, cela-, ç’ek’o-, ç’eşk’a-, do-, dolo-, e-, ek’o-, ela-, eşk’a-, ets’o-, eyo-, gama-, go-, gola-, goyo-, k’ok’o-, k’oşk’a-, me-, mela-, menda-, meşk’a-, meyo-, mo-, mola-, mok’o-, moşk’a-, mots’o-, moyo-, ok’o-, exo-, k’ots’o-, oxo-, gela-, k’ots’a-	Sole or agentive argument, 1st person (S-A.1): v-,p-, p’-, b, (f-) S-A.2: Ø S-A.3: Ø Patient or dative-marked argument, 1st person (P-D.1): m- P-D.2: g-, k-, k’ P-D.3:Ø	i/u-, i-, a-, o-	-t’ax- ‘break’

Table 7. Post-Root Complex-1

Root 0	Augmented stem formant 1	Causative suffix for intransitives 2	Causative suffix for transitives 3	Cusative suffix for present perfect construction 4
-t’ax- ‘break’	-am	-in	-ap	-ap

Table 8. Post-Root Complex-2

Thematic suffix 5	Imperfect stem formant 6	Subjunctive marker 7	Person suffixes 8
-am,-um, -e(r),-u(r)	-t’	-a	Past (PST) S-A.1.PST: -i S-A.2.PST: -i S-A.3.PST: -u Present (PRS) S-A.1.PRS: Ø S-A.2.PRS: Ø S-A.3.PRS: -s,-n

Table 9. Post-Root Complex-3

Conditional marker 9	Plurality 10	Auxiliaries 11
-k’o	-t (1st & 2nd) ; -es, -(a)n (3rd)	-(e)re, -ert’u

Here I will explain the functions of each of these affixal positions, and also, focus on morphotactic constraints and long-distance dependencies observed inside the verbal complex since this information is crucial for defining lexical rules as FSTs in order to correctly analyze word forms and produce correct word forms. However, note that most of the verbal affixes such as preverbs, valency-related vowels, causative suffixes, and thematic suffixes depend on the semantics of the verb stem related but not limited to the argument structure, transitivity, and verb class. This information on the verbal stems in Laz is something I do not have access to at the moment due to a lack of resources. Therefore, the morphotactic constraints I defined in the FST, unfortunately do not include this information which unsurprisingly leads to the *overgeneration* of word forms that are not found in the natural language but has less adverse effect on the performance of the FST as recognizer/analyzer.

3.5.1.2.1 Affirmative Preverbs

Position -4 is occupied by affirmative preverbs (PRV). They are used to increase the *certainty* of the action or the event expressed by the predicate. Only the *ko-* preverb can be followed by spatial preverbs. They are not obligatory but when they are realized on the verb stem, negative markers cannot be used because of semantic incompatibility. Additionally, the selection between these preverbs depends on the verb stem.

3.5.1.2.2 Spatial Preverbs

Spatial preverbs (PRV) occur at position -3. They basically add some kind of *spatial* meaning such as directionality or location to the verb stem which denotes motion or action itself seen in 38 (Eren, 2016). However, the non-action verbs in 39 exhibit idiosyncratic meanings after the combination of preverbs with the verb stem. Since regardless of their meaning, all spatial preverbs are subject to morphophonological changes depending on the following sounds and there can be person prefixes or version vowels surfacing between preverbs and the verb root, the verbs such as those given in 39 are still analyzed into their morphemes instead of being treated as one morphological unit for the morphological analyzer. See Holisky (1991), Öztürk and Pöchtrager (2011), Kutscher (2011), Eren (2016), and Bucak'lişi and Kojima (2003) for a more extensive account on the syntax and semantics of spatial preverbs in Laz. The examples in 38 and 39 are all taken from Öztürk and Pöchtrager (2011) and Eren (2016).

(38) Regular spatial preverbs

ce-	*'down'*	*ce-xt'-u*	*'She/he went down.'*
gama-	'out'	gama-xt'-u	'She/he went out.'

(39) Lexicalized spatial preverbs

do-guru	*'to learn'*
ce-k'itxu	'to ask again'
gama-k'otu	'to slap'

3.5.1.2.3 Person markings: Person prefixes and person suffixes

The position -2 belongs to person prefixes which are used to provide the person information of not only the subject but also other syntactic elements such as objects and if available, DAT marked applied arguments which are defined in Section

3.5.1.1.2. However, person prefixes do not reflect the number information unlike some person suffixes at position 8. In the literature, two sets of person prefixes are defined for this position as *v*-set for marking sole or agentive arguments (usually subjects) represented as S-A, and *m*-set for marking patient or dative-marked arguments (usually objects and applied arguments) which is abbreviated as P-D for the rest of the thesis. Both sets are given in Tables 10 and 11, respectively. Whereas the *v*-set is considered to reflect a *joint* agreement between the person value of the person prefix at position -2 and the person suffix at position 8, *m*-set reflects a *disjoint* agreement (see Demirok (2013) for details). This means that there are dependencies and constraints between preverbal (-2) and postverbal (8) person markers[35], also introduced in Tables 10 and 12.

Position 8 marks only the person value of the subject (S-A argument) unlike position -2. Therefore, in terms of *joint* agreement, if a person prefix referring to a S-A arguments is used preverbally in -2, it should have the same person value of the person suffix in 8. In other words, the selection among S-A person prefixes depends on the person value marked at position 8. See the example in 40 showing the joint person marking reflecting the person value of the subject *ma* ‘I’ (i.e., the 1st person) in both preverbal and postverbal positions. The sentence in 41 exemplifies disjoint agreement between the person prefix marking 1st person (*ma* ‘I’) in -2 and the person suffix marking 3rd person (*bere* ‘child’) at position 8.

(40) *Ma v-i-nçir-i.*
I.ERG S-A.1-VAL-swim-PST.1SG
‘I swam.’

Source: adapted from Öztürk and Pöchtrager (2011)

(41) Applicative construction: Location

Bere-k ma m-a-nçir-u.
Child-ERG I.DAT D-P.1-VAL:APPL-swim-PST.3SG

‘The child swam around me.’

[35] I leave the discussion on tense and number marking after the discussion over joint and disjoint markers.

At this point I will define a very important morphotactic constraint for the realization of *joint* person markers.

Morphotactic constraint 3: If preverbal (-2) S-A person markers are realized on the verbal complex, they need to reflect the same person value as the postverbal (8) S-A person markers.

For example, if the person suffix is realized as *i-* and marked as S-A.1 denoting the values of 1st person and past tense, position -2 cannot be left empty and marked with neither S-A.2 nor S-A.3[36] which would contradict with the 1st person subject suffix. The corresponding person prefix should reflect the 1st person subject and be overtly marked with the *v*-set.

Table 10. Person Markers for S-A Arguments in *Joint* Agreement

	Person prefix (S-A) -2	Root	Person suffix (only S-A) 8	
			Past	Present
1	*v-, p-, p'-, b-, f-*	Root	*-i*	Ø
2	Ø		*-i*	Ø
3	Ø		*-u* (SG), *-es* (PL)	*-s, -n* (SG), *-(a)n* (PL)

Table 11. Person Markers for P-D Arguments

	Person prefix (P-D) -2
1	*m-*
2	*g-, k, k'*
3	Ø

As an additional note on joint person markers in terms of morphosyntax and semantics, the *v*-set or S-A person prefixes *usually* mark the subject unless the subject is DAT marked (semantically the deagentive or abilitative argument) since the person

[36] I defined the zero exponents such as S-A.2, S-A.3, and P-D.3 person prefixes in the FST. Therefore, every possible case or combination shown in Tables 10 and 12 is formally defined in the analyzer. This means that an output label for an unrealized prefix produced by the analyzer is intentional and not by chance due to defined dependencies between person prefixes and suffixes.

information of a DAT subject can never surface at position 8 and thus never at position -2 realized as S-A person prefixes (Demirok, 2013), making these types of constructions (i.e., inversion) not available for *joint* person markings.

See the inflection on the verb ‘beat’ in 42 where the person suffix has the 3rd person value *by default* instead of reflecting the person information of the subject *ma* ‘I’[37]. The verbal inflection seen in 43, however, shows that when there is no inversion and necessarily a DAT subject, the postverbal position is available for marking the ERG subject. The DAT marked subjects under inversion operations shown in 42 are marked on the verb only as preverbally with P-D person prefixes (*m*-set) but not with S-A person prefixes (*v*-set).

(42) Inversion: Ability and involuntary constructions

Ma	*si*	*ce-m-a-ç-u.*
I.DAT	you.NOM	PRV-P-D.1-VAL:APPL-beat-PST.3SG

‘I was able to beat you.’/‘I involuntarily beat you.’

Source: Öztürk (2021)

(43) No inversion

Ma	*si*	*ce-k-ç-i.*
I.ERG	you.DAT	PRV-P-D.2-beat-PST.1SG

‘I beat you.’

Source: Öztürk (2021)

DAT subjects are marked with the *m*-set or P-D person prefixes in the verbal complex. However, when focused, the NOM object can mark position -2 and on the surface act as the subject (due to the mechanism behind inversion constructions (Öztürk, 2013)). Therefore, the NOM object (semantically the patient/theme argument) can also be marked preverbally with the *v*-set or S-A person prefixes resulting in joint person marking. See 44 below.

[37] Due to the DAT case marking, the deagentive or abilitative subject is marked with P-D person prefixes only realized preverbally. Also, note that in 40 and 41, *ma* refers to both DAT and ERG marked 1st person pronoun. Pazar Laz does not differentiate between 1st and 2sn person DAT, ERG, and NOM (Öztürk & Pöchtrager, 2011).

(44) Inversion with the focused NOM object

Himu-s *ma* *ce-v-a-ç-i.*
He/she-DAT I.NOM PRV-S-A.1-VAL:APPL-beat-PST.1SG

'He/she was able to beat ME.'

Source: Öztürk and Pöchtrager (2011)

Therefore, a possible morphotactic constraint is in fact eliminated here. The default 3rd person suffix and use of only D-P argument person prefixes by ignoring any joint S-A person prefix could be defined as a morphotactic rule in relation to the inversion constructions. However, with focused objects, the S-A person prefixes are made possible for these constructions so no constraint is needed at least for the person markings, but there is still a relationship between the valency-related vowels and the thematic suffixes for these cases which will be discussed later in Section 3.5.1.2.4.

In the case of *disjoint* person markers on the verbal complex, the person prefix (-2) reflects the person information of a P-D argument, and it cannot match with the person value of the subject realized post-verbally (8) seen in the sentences in 41 and 43 before. However, the only exception to this is the 3rd person markers since the subject and the object can refer to different 3rd person individuals in a discourse. However, it is not acceptable to use the (*m*-set) person prefix denoting the 1st person value of a P-D argument while the person suffix also reflects the 1st person value, because this would mean that both the S-A argument (mostly the subject) and a D-P argument (mostly the object) are the same [38].

Table 12 defines the possible combinations between P-D person prefixes and person suffixes[39]. I already explained why P-D.1 + A-S.1 and P-D.2 + A-S.2 don't

[38] When the subject and the object refer to the same person, the structure goes under reflexivization and gets the related version vowel *i-* seen in (i) below. The person prefix is also from *v*-set and not from *m*-set referring to the S-A argument.

(i) Reflexivization

Ma *yali-s* *v-i-dzir-am-Ø.*
I.NOM mirros-DAT S-A.1-VAL:REFL-see-TS-PRS.1SG

'I see myself in the mirror.'
Source: Öztürk and Pöchtrager (2011)

[39] Table 12 only shows the inflections with person suffixes having past tense value, but not non-past

work and P-D.3 + A-S.3 works in terms of morphotactics. However, the reason why the combination of P-D.3 prefix and S-A.1 suffix is not possible is not due to disjoint agreement since they do not share the same person value but the *hierarchy* between person prefixes for the position -2. As mentioned before, preverbal markers ‘compete for limited morphological slots’(Atlamaz, 2013; Öztürk, 2013).

Table 12. Person Markers for P-D Arguments in *Disjoint* Agreement

Person prefix (P-D) -2				
1	2	3		
N/A	*g, k, k'*- root -*i*	N/A	1	Past person suffix (A-S) 8
m- root -*i*	N/A	Ø- root -*i*	2	
m- root -*u*	*g, k, k'*- root -*u*	Ø- root -*u*	3	

There is a hierarchy (or a *competition* (Atlamaz, 2013)) between arguments to fill the position -2 depending on their case markings and semantic roles. This hierarchy is formalized as 45 according to Öztürk and Pöchtrager (2011) and as 46 based on case markings according to Öztürk (2013). Morphotactic effect of this competition between arguments is observed that a specific person value (i.e., P-D.3) cannot be overtly marked in this position, and in Pazar Laz, this position is *always* filled if there is an available argument in the structure that could overtly mark this position[40].

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

Source: Öztürk and Pöchtrager (2011)

(46) DAT1/2 > NOM1/2 > ERG1/2 > DAT3=ERG3=NOM3

Source: Öztürk (2013)

since the tense information does not affect the possible combinations defined in the table. The same morphotactics are used for the non-past in the FST.

[40] However, there could be more than one available argument and, in this case, the hierarchy applies for choosing the argument that will mark the position -2. Since the morphological analyzer cannot interpret the syntactic environment of the verb (e.g., whether or not there is a patient argument or the dative-marked applied argument in the structure, or how many arguments there are) but have access to word-level information only, this hierarchy matters in terms of how it affects which combinations between person prefixes and affixes are possible or not at the word-level.

Notice that the first three levels defined in 45 (D1/2 > P1/2 > A1) point to overtly marked person prefixes while the last level (D3=P3=A2/3) consists of arguments that cannot be overtly marked (see Tables 10 and 11). This means that position -2 is or can be left empty as the last choice. However, under one condition, it is never empty. If position 8 is marked with S-A.1 which in turn makes the S-A.1 prefix available for position -2, then position -2 will be always filled since S-A.1 prefix is overt (*v*-set). In other words, zero exponent P-D.3 person prefix (at the last level) will be disallowed to mark this position since the S-A.1 argument is higher in the hierarchy and has an overt marking. This constraint is defined by marking the combination between P-D.3 and S-A.1 as N/A in Table 12. These constraints combined in Table 12 are defined as two morphotactic constraints;

Morphotactic constraint 4: Disjoint person markings cannot share the same person value except 3rd person.

Morphotactic constraint 5: If position 8 is marked with 1st person, position -2 is never empty. In this case, the unmarked P-D.3 prefix is disallowed before the S-A.1 suffix.

Other issues related to person markers are the tense and number information. Pazar Laz has two sets of person suffixes to differentiate between present and past tense which are shown in Table 10. Plurality is marked with *-t* when the person suffix denotes the first or the second person regardless of tense. Pazar Laz fuses plurality with the third person suffixes as *-es* and *-(a)n*, marking past and present tense respectively. Also, note that Laz expresses plurality at position 10 which makes the plural form depending on the person suffix although the plurality can belong to the prefixally marked argument. Therefore, this creates ambiguity in the case of a *disjoint* person marking without an analysis on the sentence level. As seen in the sentence in 47, the plural marking is due to the prefixally marked D-P argument *şk'u* 'us' and not the singular S-A argument *cumadi* 'uncle'. However, without the surrounding overt arguments in the sentence, on the word level, the predicate is ambiguous which potentially leads to the readings of the S-A argument showing plurality (i.e., *cumadepe* ‘uncles’) and the D-P argument showing singularity (i.e., *ma* ‘me’) or both S-A and D-A arguments being plural.

(47) *Cumadi-k* *şk’u* *opşa* *m-o-dits-in-es.*
uncle-ERG we.NOM very D-P.1-VAL:CAUS-laugh-CAUS-PST.3PL
‘(Some) uncle made us laugh so much.’

Source: Bucak’lişi et al. (2007)

Morphotactic constraint 6: If the person suffix denotes 1st or 2nd person, the plural marking *-t* should follow the person suffix. Plurality with 3rd person only allows fused forms and prevents marking of the singular counterparts given in Table 10 before plural denoting fused suffixes.

3.5.1.2.4 Valency-related vowels and operations

Valency-related vowels or pre-root version vowels (i.e., *i-*, *u-*, *a-*, *o*) directly attach to the verb stem at position -1. They signal changes to the valency or the argument structure of the predicate resulting from applicativization (*i-/u-*, *a-*), causativization (*o-*), and inversion (*i-/u-*, *a-*) phenomena. They are mostly considered to manifest as part of applicative morphology except for causativization. There is also the pronominal clitic *i-* used for reflexivization, reciprocals, and active impersonal (voice) constructions (AIC) (i.e., passive, impersonal passive, and anticausative) to be able to reflect the suppressed internal or external argument.

Valency-related vowels can also be *inherent* or lexical depending on the semantics of the verb stem such as those in 48 (Öztürk & Pöchtrager, 2011). Öztürk (2021) discusses the *clitic* status of the valency vowel *i-* on the verbs in 48. According to Öztürk (2021), *i-* found in ‘intransitive’ unergatives such as *-bir-* ‘play’ in 48 indicates the presence of an underlying transitive verb. It works as a pronominal clitic filling the semantic role of *undergoer/patient* and obligatorily refers to (or co-indexed with) the subject (the initiator) similar to reflexivization.

Some transitive verbs with inherent vowels are in fact considered to be applicativized since there is a DAT marked applied argument in the structure and also applicative morphology observed on the verb stem (e.g., psychology verbs or psych-applicatives as inversion constructions usually marked with *a-* and in some cases *o-* seen in 49 and 50 respectively) (Demirok, 2013; Öztürk, 2013). Although it is their inherent properties that require them to be inflected with applicative morphology, I will consider them under inversion constructions and thus as applicatives (i.e., *Appl: Psych*, see Table 14) in accord with the literature. Additional examples for inherent vowels are given in 51 and 52.

(48) Inherent valency-related vowels (*i-* and *u-*)

Ma v-i-bir/bgar/nçir/çaliş-i / v-u-k’ap’-i.
I.ERG S-A.1-VAL-play/cry/swim/work-PST.1SG S-A.1-VAL-run-PST.1SG

‘I played/cried/swam/worked / ran.’)

(49) Inherent valency-related vowel *a-*

Bere-s bozomota a-limb-e-n
child-DAT girl.NOM VAL:APPL-love-TS-PRS.3SG

‘The child loves the girl.’

(50) Inherent valency-related vowel *o-*

Ma si go-m-o-ç’ondr-u.
I.DAT you.NOM PRV-P-D.1-VAL:APPL-forgot-PST.3SG

‘I forgot you.’

(51) Inherent valency-related vowel *a-*

Bere-k bozomota-s gv-a-xel-u.
child-ERG girl-DAT PRV-VAL-kiss-PST.3SG

‘The child kissed the girl.’

(52) Inherent valency-related vowel *o-*

Bere-k ma gza m-o-ts’ir-am-s.
child-ERG I.DAT road/way.NOM D-P.1-VAL-show-PRS.3SG

‘The child is showing me the way.’

Since there is no source at the moment with which I can categorize verbs based on their semantics and argument structure inside the lexicon, and thus which inherent vowel they take, it is beyond my capacity to assign *inherent* valency-related vowels for each such verb. Therefore, there is no constraint over inherent vowels. They inflect every verb stem in the lexicon labeled as VAL.

Interestingly, these valency-related vowels never appear in the nominal or participle form of verbs (e.g., *o-bir-u*[41] 'playing') (Öztürk, 2021) which can also be observed in the last two forms of -*mbon*- 'wash' in 53. Therefore, although they are *inherent*, they don't inflect the word stem in every case, for any word form. They are replaced by applicative vowels when there is a valency-changing operation. The inversion construction adding the abilitive modality changes *i*- to *a*- as seen in 53 (see the non-applicative form of the verb -*bir*- in 48). As a result, they are treated as optional inflectional prefixes.

(53) *Ma* *m-a-bir-i*
I.DAT S-A.1-VAL:APPL-play-TS-PST.3SG
'I was able to play.'

Source: adapted from Bucak'lişi et al. (2007)

Valency-related vowels are introduced in 54 with their semantics and corresponding operations respectively. See footnote 41 for the explanation on the abbreviations.

(54) The change of valency-related vowels (italicized) on the verb -*mbon*-'wash'

mbon-um-s	'He/she is washing smt.'	
u-mbon-am-s	'He/she is washing smo else.'	Appl.: Benef.
i-mbon-am-s	'He/she is washing a part of self.'/ 'He/she is washing (him/herself).'	Refl.
i-mbon-e-n	'Smt get washed'	AIC
a-mbon-e-n	'He/she can/is able to wash smt.'	Inver.: Abil.
u-mbon-ap-u-n	'He/she has washed smt before.'	Inver.: Exper. perf.
o-mbon-ap-am-s	'He/she is making smo wash smt.'	Caus.
o-mbon-u	'Washing (action)'	Nomnlz.
mbon-eri xe	'Washed hand'	Part.

Source: Bucak'lişi et al. (2007)

[41] *o*- is not a valency-related vowel but part of a cimcumfix *o*-...-*u* which derives verbal nouns.

Note: Abbreviations are as follows; smt=something; smo=someone; Nmnlz.=Nominalization; Part.=Participle; Appl.: Benef.=Applicativization: Benefactor; Refl.=Reflective; Impr. pass.=Impersonal passive; Inver.: Abil.=Inversion: Abilitative; Inver.: Exper. perf.=Inversion: Experiential perfect; Caus.=Causativization

Now, I will define morphotactic rules for these constructions. I will use valency-related vowels as 'anchors' for the corresponding constructions and define long-term dependencies with person markings as well as thematic and causative suffixes. Later, I will explain how these constructions can be combined in a clause, and how the valency vowel changes (or *overwritten* (Öztürk & Pöchtrager, 2011)) on the verb accordingly since only one can occupy position -1. This situation is, in fact, discussed as a kind of *competition* between valency-related vowels over position -1 depending on the hierarchical level of the applicative constructions discussed in Öztürk (2013) as low, high, and higher, and also c-commanding relations.

There are several valency-changing operations in Pazar Laz. I divided and listed them based on their shared morphology (i.e., valency-related vowels) as well as the semantics they add to the predicate given in Tables 13, 14, and 15. Those marked with (*) and (**) show long-distance dependencies with thematic suffixes, *-e(r)* and *-u(r)* respectively. Additionally, (***) marks dependencies with causative suffixes realizing at three consecutive positions in the verbal complex (i.e., 2, 3, and 4).

Table 13. Valency-Increasing Operations

i-/u-	*a-*	*o-*
Appl: Recipient Appl: Possessor Appl: Benefactor	Appl: Location	Causative***

There is more to these operations than just adding a valency-related vowel. Those that require the *i-/u-* valency-related vowels (see Table 13 and 14) choose between these two vowels based on the person information of the DAT marked applied

Table 14. Inversion (Modality)

i-/u-	*a-*
Appl: Experiential perfect**/***	Appl: Abilitative* Appl: Deagentive* Appl: Psych*

Table 15. Pronominal Cliticization

i-
AIC (Passive, impersonal passive*** and anticausative)* Reflexive Reciprocal

argument[42]. *i-* is used when the applied argument is 1st person (*m*-set) or 2nd person (*g-, k, k'*) while *u-* is used for 3rd person applied arguments[43] seen in the examples in 55, 56, and 57, respectively.

(55) *Si* *ma* *past'a* *m-i-ç'v-i*
you.ERG I.DAT cake.NOM D-P.1-VAL:APPL-play-PST.2SG
'You baked me a cake.'

(56) *Ma* *si* *past'a* *g-i-ç'v-i*
I.ERG cake.NOM D-P.2-VAL:APPL-play-PST.1SG
'I baked you a cake.'

Source: Öztürk (2013)

(57) *Ma* *bere-s* *past'a* *v-u-ç'v-i*
I.ERG child-DAT cake.NOM S-A.1-VAL:APPL-bake-PST.1SG
'I baked the child a cake.'

Source: Öztürk and Pöchtrager (2011)

[42] The term *applied argument* is used to refer to both a non-core argument added into the argument structure and the suppressed subject under an inversion operation (i.e., present perfect construction) as an umbrella term. Applied arguments are *always* DAT marked.

[43] Since P-D.3 is unmarked, the hierarchy applies for the person prefix position and the S-A.1 person marking fills the position since there is no other available argument, and the position cannot be left empty. However, this does not affect the semantics of *u-* vowel as reflecting the 3rd person benefactor (i.e., *bere* 'child'). Refer to 30 to see the verbal inflection when the subject is 3rd person; therefore, the preverbal position is left empty.

I will not provide examples of other valency-increasing operations that are also constructed with *i-/u-* vowels since applicative morphology reflected on the verbal complex, is the same. See Öztürk and Pöchtrager (2011) and Öztürk (2013) for more examples on these constructions. However, note that *Appl: Experiential perfect* as an inversion modality construction requires the verb to be marked with both a causative suffix, specifically *-ap* at position 4, and a thematic suffix, specifically *-u(r)* at position 5 as seen in 58.

(58) *Ma* *past'a* *u-ç'v-ap-u-n*
I.DAT cake.NOM VAL:APPL-bake-CAUS-TS-PRS.3SG
'I have baked a cake.'

Now, I will define a morphotactic rule to restrict the unacceptable combinations between the valency-related vowels *i-/u-* and person prefixes.

Morphotactic constraint 7: In case of applicatives such as *Appl: Recipient, Appl: Possessor, Appl: Benefactor, Appl: Experiential perfect*, if the *i-* vowel fills the position -1, it is always preceded by either the P-D.1 person prefix (*m-*) or a P-D.2 person prefix (*g, k, k'*), and if *u-* is used, the person prefix can be unmarked reflecting the value of D-P.3 or can be in joint agreement with the person prefix at position 8, and overtly marked only if the person prefix is 1st person[44].

Morphotactic constraint 8: Under *Appl: Experiential perfect* construction, following the vowel *i-/u-* the verb complex needs to be marked with the causative suffix *-ap* and the thematic suffix *-u(r).*

The verbal inflectional pattern for the applicatives using the vowel *a-*, namely *Appl: Location* as the valency-increasing applicative and *Appl: Abilitative, Appl: Deagentive* and *Appl: Psych* as inversion applicatives does not show any morphotactic constraints related to person prefixes. There is no alternation for the valency vowel *a-* unlike the *i-/u-* case. Any person prefix can come before *a-*. Therefore, I did not put an additional constraint on person prefixes for *a-* for these

[44] With these instructions, the analyzer will give 3 different outputs when the person prefix is unmarked preceding the *u-* vowel. The hierarchical structure between person prefixes drawn in 45 does not differentiate between unrealized D-P.3, S-A.2 and S-A.3 either. Theoretically, all these three cases are possible with the *u-* vowel.

applicatives. However, note that for inversion operations (i.e., *Appl: Abilitative, Appl: Psych, Appl: Experiential perfect (i-/u-)*), I consider cases where the NOM object is focused; therefore, the joint person markings are also possible[45]. An example for the focused NOM object (capitalized) is provided in 59 in relation with sentences given in 42 (showing the inversion operation without focus) and 43 (showing the no inversion case). Examples for *Appl: Location* are also given in 60.

(59) *Ma si ce-m-a-ç-i*
I.ERG you.NOM PRV-D-P.1-VAL:APPL-beat-PST.2SG
'I was able to beat YOU.'

Source: adapted from Öztürk and Pöchtrager (2011)

(60) *Ma bere-s v-a-nçir-i*
I.ERG child-DAT S-A.1-VAL:APPL-swim-PST.1SG
'I swam around the child.'

Source: Öztürk and Pöchtrager (2011)

Morphotactic constraint 9: The verb stem is marked with the valency related vowel *a-* under the constructions as *Appl: Location*, *Appl: Abilitative*, *Appl: Deagentive* and *Appl: Psych*. While *Appl: Abilitative*, *Appl: Deagentive* and *Appl: Psych* also require the thematic suffix *-e(r)* to express the backgrounded agentive subject, *Appl: Location* does not need any thematic suffix to be able to surface.

Valency-related vowel *o-* is used for causatives and it does not alternate in form and does not bear any dependencies with the person markers. Therefore, there is no morphotactic constraint for causative constructions related to person markers. However, causatives require causative suffixes at position 2 or at position 3 or both depending on the transitivity of the predicate. Intransitives take *-in* at position 2 while transitive take *-ap* at position 3 under causative constructions (see 63 and 62 respectively). In the case of *double* causative, the intransitive verb base is marked with both *-in* and *-ap* at the same time as seen in 64. All examples given in 61-64 are taken and adapted from Bucak'lişi et al. (2007). Also, note that intransitive bases when causativized introduce ERG causer whereas transitive bases introduce DAT

[45] Also see the related discussion on the morphosyntax of these constructions in Section 3.5.1.2.3.

marked causee in addition to ERG causer (see examples in 64 and 62 respectively). This is important because when they are combined with other valency-changing constructions, the level and case marking of arguments will determine both person marking and the valency-related vowel on the verb complex (Demirok, 2013).

(61) Non-causative (transitive)

Bere-k mjalva nts'or-um-s.
girl-ERG milk.NOM filter-TS-PRS.3SG

'The child filtered the milk.'

(62) Causative (transitive)

Baba-k bere-s mjalva
father-ERG child-DAT milk.NOM
o-nts'or-ap-am-s.
VAL:CAUS-filter-$\text{CAUS}_{\text{transitive}}$ – TS – PRS.3SG

'The father made the child filtered the milk.'

(63) Non-causative (intransitive)

Nceni i-mord-u.
calf.NOM VAL-grow-PST.3SG

'The calf grew.'

(64) Causative (intransitive)

Nana-k nceni o-mord-in-am-s.
mother-ERG calf.NOM VAL:CAUS-grow-$\text{CAUS}_{\text{intranstitive}}$ – TS – PRS.3SG

'The mother is raising the calf.'

(65) Double causative (intransitive)

Nana-k baba-s nceni
mother-ERG father-DAT calf.NOM
o-mord-in-ap-am-s
VAL:CAUS-grow-CAUS- CAUS-TS-PRS.3SG

'The mother is making the father raise the calf.'

Morphotactic constraint 10: When causativized and marked with *o-*, the verb is required to be inflected with causative suffixes as *-in* and *-ap.* At least one of them should be present for the verb stem to be able to express the causative.

The valency-related vowel *i-* used as the pronominal clitic for so called valency-decreasing operations given in Table 15. Öztürk (2021) argues that these operations cause the overt internal (i.e., object) or external (i.e., subject) arguments to be eliminated from the sentence although the inflection of *i-* as the pronominal clitic in the verbal complex still preserves these arguments in the structure. This is also discussed as being similar to the inherent pronominal clitic *i-* in transitive unergatives such as *-bir-* 'play' discussed before. Therefore, note that these are not truly valency-decreasing operations although some literature refers to them as such.

With reflexives, it is the internal argument that is co-indexed with the subject in terms of person value is reflected into this vowel *i-* and on the syntactic level the overt form that keeps the person information of this internal argument cannot be used (see 67) (Öztürk, 2021). However, if the sentence has the overt reflexive pronoun such as *çendi* in 67, then the valency-related vowel/pronominal clitic *i-* cannot surface, similar to the non-reflexive given in 68. Therefore, the pronominal clitic *i-* and reflexive pronouns are mutually exclusive. In this case, the pronominal clitic introduces and carries the semantic role *undergoer/patient.*

(66) Reflexive (pronominal clitic)

Ma yali-s v-i-dzir-am-Ø.
I.ERG mirror-DAT S-A.1-VAL:REFL-see-TS-PRS.1SG
'I see myself in the mirror.' Literal translation: 'I appear in the mirror.'

(67) Reflexive (reflexive pronoun)

Ma yali-s çendi b-dzir-am-Ø.
I.ERG mirror.DAT self.NOM S-A.1-see-TS-PRS.1SG
'I see myself in the mirror.'

(68) Non-reflexive

Ma yali-s si g-dzir-am-Ø.
I.ERG mirror.DAT you.NOM S-A.2-see-TS-PRS.1SG

'I see you in the mirror.'

Other types of reflexives are similar to valency-increasing applicative constructions (Öztürk & Pöchtrager, 2011). They introduce similar semantic roles into the structure such as possessor and benefactor as seen in 69 and 70 respectively. However, in reflexives, the person value of the subject should be reflected onto these new roles marked as the pronominal clitic *i-*; therefore, there is no overt argument in the sentence corresponding to these roles whereas in applicatives the applied arguments bearing the roles are overt and DAT marked, and it does not reflect the person value of the subject seen in 71.

(69) Reflexive introducing possessor

Bere-k k'iti i-ts'uts'on-am-s
child.ERG thumb.NOM VAL:REFL-suck-TS-PRS.3SG

'The child is sucking his/her (own) thumb.'

(70) Reflexive introducing benefactor

Nana-k uşkuri i-ts'il-am-s
mother-ERG fruit.NOM VAL:REFL-pick-TS-PRS.3SG

'The mother is picking fruit for herself.'

(71) Applicative introducing benefactor

Evro-k ma uşkuri m-i-ts'il-am-s
Evro-ERG I.DAT fruit.NOM D-P.1-VAL:APPL-pick-TS-PRS.3SG

'Evro is picking fruit for me.'

In terms of person marking, since both the internal and the external argument are co-indexed, the joint person agreement is allowed for reflexives. Therefore, I will not put a morphotactic constraint on person markings for reflexives. This is the case for reciprocals as well with the addition of the preverb *ok'o-* (see 70).

(72) Reciprocal

Sk'u *ok'o-v-i-dzir-i-t*
we.NOM PRV-S-A.1-VAL:REFL-see-PST.2SG-PL

'We saw each other.'

Morphotactic constraint 11: Under pronominal cliticization, the verb is marked with *i*- and the person markers should be in joint agreement.

Morphotactic constraint 12: Realization of the reciprocal *i*- is dependent on the spatial preverb *ok'o-*.

Other pronominal cliticization constructions discussed in the literature such as impersonal passives, passives and anticausatives are grouped under the term *active* impersonal construction (AIC) (Öztürk, 2021). AIC is defined as the 'umbrella' construction covering all these readings that have 'strictly externally caused agentive reading referring to a human agent'. Different from reflexives, the pronominal clitic *i*- this time reflects the external argument (i.e., the subject) instead of the internal argument (i.e., the object). When being compared to 72, 73 shows that pronominal clitic *i*- reflects the suppressed external argument (*the contractor*) for the transitive verb *-rg-* 'build'. AICs can be also used with unergatives and unaccusatives and turn them into impersonal passives with the addition of *-in* causative marker. It can be also observed that AIC requires the *-e(r)* thematic suffix (see 74 and 75).

(73) *Ust'a-k* *oxori* *mo-rg-am-s.*
contractor.ERG house.NOM PRV-build-TS-PRS.3SG
'The contractor is building the house.'

Source: Bucak'lişi et al. (2007)

(74) Active impersonal: Anticausative / passive

oxori *mv-i-rg-e-n.*
house.NOM PRV-VAL:AIC-bring-TS-PRS.3SG

'The house gets built.' / 'The house is being built.' (Someone is building the house)

Source: Öztürk (2021)

(75) Active impersonal: Impersonal passive

i-xap'ar-in-e-n
VAL:AIC-talk-CAUS-TS-PRS.3SG

'One can talk.' (People are talking):

Source: Bucak'lişi et al. (2007)

Morphotactic constraint 13: Under AIC operations, the verb is marked with the pronominal clitic *i-* and the thematic suffix *-e(r)*. In case of unergatives and unaccusatives, it is also marked with the causative marker *-in.* Since there isn't a list of unergatives and unaccusatives available, the *-in* marks every verb stem optionally.

There is another AIC construction denoting impersonal passive with an applicative benefactor argument (Demirok, 2013; Öztürk, 2021). The construction is formed with *a-* as the valency-related vowel and *-e(r)* as the thematic suffix. It has the same morphology as the inversion constructions marked with *a-*. Therefore, it is in fact ambiguous and leads to two different readings as seen in 76. The APPL only reading in 76 has the DAT marked *xordza* 'woman' as the external argument, and the APPL + AIC reading has it as the applied benefactive argument. I also found the reading with the possessor argument as given in 77 from the dictionary. Note that this construction is not the surface combination of *Appl: Benefactor* and *AIC: Impersonal passive* since *Appl: Benefactor* is marked with the valency-related vowel *i-/u-* only and *AIC: Impersonal passive* is marked with the impersonal clitic *i-* and the thematic suffix *-e(r)* (for transitive verb stems). Neither of these constructions have the valency-related vowel *a-*. Therefore, it needs to be classified separately and *a-* should have another label as VAL:AIC in addition to VAL:APPL defined for only applicatives.

(76) *Oxori* *xordza-s* *mv-a-rg-e-n.*
house.NOM woman.DAT PRV-VAL:APPL/VAL:AIC-build-TS-PRS.3SG
APPL: 'The woman is able to build the house.'

APPL + AIC: 'The house is being built for the woman.'

Source: Öztürk (2021)

(77) *Oxori* *dolv-a-xv-es*
house.NOM PRV-VAL:AIC-demolish-PST.3PL

APPL + AIC: ‘Their house was demolished.’

Source: Bucak’lişi et al. (2007)

Morphotactic constraint 14: The AIC construction denoting impersonal passive with an applicative benefactor argument requires the verb to be marked with the valency-related vowel *a*- and the thematic suffix *-e(r)*.

Up until this point, I have defined several morphotactic rules on valency vowels denoting different constructions. I tried to label these vowels in accord with the literature related to Pazar Laz grammar as seen in Table 16 although my objection here is to define the morphotactics and individual dependencies for each operation correctly as much as possible rather than focusing on labelling too much. Note that one could always change the labels and the naming for the morphemes according to the use case of the analyzer, and if needed for a specific task, a more detailed labelling or a more general one could be chosen and applied (e.g., Universal Dependencies feature tag set).

I will look at how combining different constructions, individually defined in Table 16 changes the verbal complex and creates word forms that cannot be covered by the morphotactics I defined so far.

Note that I used valency-related vowels as *anchors* for defining morphotactics which means that any other affix in the construction will surface only when the corresponding valency-related vowel realizes at position -1. For example, the causative marker -*ap* at position 3 (refer to Table 16) will be marked on the verb only if the valency-related vowel is selected as *o*- labelled with VAL:CAUS, and similarly, the thematic suffix *-e(r)* can be used only if the vowel *a*- as VAL:APPL or VAL:AIC, or the vowel *i*- as VAL:AIC occupies the position -1[46].

The sentence in 78 exemplifies the combination of experiential perfect applicative and causative. It can be seen that the inflection on the verbal complex is

[46] The *-e(r)* suffix is highly predictable in Pazar Laz since it can only be found in applicatives or AICs signaling the backgrounded agent subject, and not elsewhere.

Table 16. Labels and Dependents of Valency-Related Vowels

	Operations	Vowel(s)	Label	Causative suffix(es)			Thematic suffix
		-1		2	3	4	5
Appl	Appl: Recipient	*i-/u-*	VAL:APPL	N/A			N/A
	Appl: Possessor	*i-/u-*	VAL:APPL				
	Appl: Benefactor	*i-/u-*	VAL:APPL				
	Appl: Location	*a-*	VAL:APPL				
	Appl: Experiential perfect	*i-/u-*	VAL:APPL			*-ap*	*-u(r)*
	Appl: Abilitative	*a-*	VAL:APPL	N/A			*-e(r)*
	Appl: Deagentive	*a-*	VAL:APPL				
	Appl: Psych	*a-*	VAL:APPL				
Caus	Caus (for intransitives)	*o-*	VAL:CAUS	*-in*			N/A
	Caus (for transitives)	*o-*	VAL:CAUS		*-ap*		
	Caus (for doubles)	*o-*	VAL:CAUS	*-in*	*-ap*		
AIC	AIC: Passive	*i-*	VAL:AIC	N/A			*-e(r)*
	AIC: Anticausative	*i-*	VAL:AIC				
	AIC: Impersonal passive	*i-*	VAL:AIC	*-in*			
	Appl: Benefactor + AIC: Impersonal Passive	*a-*	VAL:AIC	N/A			
Refl	Reflexive	*i-*	VAL:REFL	N/A			N/A
Recp	Reciprocal	*i-*	VAL:RECP	N/A			N/A

not covered by any applicative morphology related to the VAL:APPL vowel *i* in Table 16. There is no morphotactics that would allow two consecutive *-ap* causative markers at the moment. Therefore, in order for the analyzer to both generate and recognize these word forms, such cases should be detected.

(78) *Ma si himu-s*
I.DAT you.DAT he/she-DAT
ce-m-i-ç-am-ap-ap-u-n
PRV-D-P.1-VAL:APPL-beat-AUG-CAUS-CAUS-TS-PRS.3SG
I have made you beat him/her before.'

Source: Öztürk and Pöchtrager (2011)

At this point, the literature needs to be visited for what kind of mechanism is behind the verbal affixation when two constructions are combined and the selection of the valency-related vowel for the position -1 when there is more than one option available. The next thing is to find a way to redefine morphotactics so that they can cover these cases.

However, there are restrictions on which and how many applicatives are allowed together per clause. It is assumed that no more than two applicatives can combine and the combination depends on the level of applicatives in the syntactic structure. A more detailed comparison between applicatives is given in Table 17, and possible and impossible combination between applicatives are given in Table 18. In general, applicatives being on the same level cannot combine (e.g., Present Perfect and Abilitative)[47].

Table 17. Levels of Applicatives in Pazar Laz

Low	High	Higher
Appl: Goal Appl: Source Appl: Possessor	Appl: Location Appl: Benefactor Appl: Psych	Appl: Deagentive Appl: Abilitative Appl: Present perfect

Source: Öztürk (2013)

Table 18. Combinations of Levels of Applicatives in Pazar Laz

	Low	High	Higher
Low	x	C	C
High	C	x	C
Higher	C	C	x

Source: Öztürk (2013)

In terms of event structure, low applicatives define a relation between two individuals in the sentence by introducing possessor, goal and source semantic roles. High applicatives on the other hand define a relation between an individual and an event by introducing benefactor/malefactor, location or experiencer (psych

[47] For a more detailed account on the combinations of different applicatives, see Öztürk (2019a).

applicatives). An additional level investigated by Öztürk (2013) to this two-level applicative approach defined by Pylkkänen (2008) is *higher* referring to what I discussed as inversion constructions that introduce experiential perfect, abilitative and deagentive *states* to an individual.

Demirok (2013) defines c-command relations between arguments that determine person agreement on the verbal complex. Öztürk (2019a) argues that the hierarchical structure formed by these c-command relations also rules the competition between valency-related vowels for the position -1. The hierarchy is defined in 79 and it can be seen that it reflects the relation between applicatives introduced in Table 17. Additionally, it shows that DAT marked causee argument introduced with causativization is higher than some applied arguments. However, these low level applicatives are only marked with valency-related vowels, and when overwritten by the causative *-o*, the verbal complex realizes purely as a causative construction on the word-level as seen in sentences in 80 and 81 unlike the situation seen in 78. Therefore, there is no need to put morphotactic rules for such combinations since these word forms are already covered by morphotactics defined for causativization.

(79) DAT (Subject)> DAT (Causee) > DAT (Benefactor/Possessor) > DAT (Goal/Possessor/Recipient) > NOM

Source: Demirok (2013)

(80) *Xordza-k ma si oşk'uri*
woman-ERG me.DAT you.DAT apple.NOM
m-o-ncğon-ap-u
D-P.1-VAL:CAUS-send-CAUS-PST.3SG
The woman sent you apple for me'

Source: Demirok (2013)

(81) *K'oçi-k ma si dişk'a m-o-çit-ap-u*
man-ERG me.DAT you.DAT wood.NOM D-P.1-VAL:CAUS-cut-CAUS-PST.3SG
The woman sent you apple for me'

Source: Demirok (2013)

Reflexives and reciprocals show the same tendency as low-level applicatives and the valency-related vowel *-i* is overwritten by the causative *-o* as seen in sentences in 82 and 83. 84 shows the uncausativized version of 83.

(82) *Nani-k bere-s o-mbon-ap-u.*
mother-ERG child-DAT VAL:CAUS-wash-CAUS-PST.3SG
'The mother made the child wash himself.'

Source: Öztürk (2021)

(83) *Gubazi-k k'oçepe-s k-ok'-o-il-ap-u*
Gubaz-ERG men-DAT PV-PV:RECIP-CAUS-fight-CAUS-PST.3SG
'Gubaz made the men fight (each other).'

Source: Bucak'lişi et al. (2007)

(84) *K'oçepe ok'-i-il-e-t'-es*
men.NOM PV:RECIP-VAL:RECIP-fight-TS-IMPF-PST.3PL
'Men was fighting (each other).'

Source: Bucak'lişi et al. (2007)

Morphotactic constraint 15: When higher applicatives are combined with causatives, the verbal complex is marked with applicative valency-related vowels while showing causative morphology (e.g., causatives suffixes -*in* and -*ap*) in addition to applicative morphology.

3.5.1.2.5 Augmentative

The position 1 refers to the augmentative (AUG) stem formant -*am* (as known as thematic suffix found at position 5). Although there is not much information found on this marker, specifically in this position, in Öztürk and Pöchtrager (2011), it is seen to be used in present perfect and future constructions similar to thematic suffixes appearing at position 5.

3.5.1.2.6 Causative Suffixes

The position 2, 3 and 4 are collectively used for causative (CAUS) markers although each position is filled according to the transitivity of the verb stem and the operation. They can appear at the same time, but they depend on the existence of certain valency-related vowels at position -1, forming long-distance dependencies.

With the applicative experiential perfect construction, the verbal complex is marked with *-ap* causative marker at position 4 after valency-related vowel *-i/u*. Other causative markers *-in* and *-ap* attaches to the verbal complex at positions 2 and 3 when causativized sometimes together with causative valency-related vowel *-o* but not with inherently causativized verb roots as seen in 85 which I found in the dictionary and also corpora. The word form in 88 shows the verb root.

(85) *Para* *baba-s* *me-v-u-nçiş-in-i*
money-NOM father-DAT PV-S-A.1-VAL-rush-CAUS-PST.1SG
'I rushed my father the money.'

Source: Bucak'lişi et al. (2007)

(86) *me-nç'iş-u*
PV-rush-PST.3SG
'He/she rushed to him/her.'

Source: Bucak'lişi et al. (2007)

3.5.1.2.7 Thematic Suffixes

Thematic suffixes (TS) follow causative suffixes holding the position 5. In terms of aspectual information, they mark the imperfective aspect on the verbal domain. In the present tense, they can have either a habitual reading or progressive reading as shown in 88 which is taken from Taylan and Öztürk (2014); however, the past tense allows only imperfective reading and the imperfective marker *-t'* should follow the thematic suffix seen in 88.

(87) *Ahmedi-k* *dişk'a* *me-ğ-am-s.*
Ahmet-ERG wood.NOM PV-bring-TS-PRS.3SG
'Ahmet brings/is bringing wood.'

Source: Taylan and Öztürk (2014)

(88) *Ahmedi-k* *dişk'a* *me-ğ-am-t'-u.*
Ahmet-ERG wood.NOM PV-bring-TS-PST.3SG
'Ahmet was bringing wood.'

Source: adapted from Taylan and Öztürk (2014)

There are four main thematic suffix in Pazar Laz also seen at position 5 in Table 8 as *-am, -um, -e(r)* and *-(u)r*. In Pazar Laz, the selection of thematic suffixes is dependent on the semantics and argument structure of the verb, and additionally, based on the *voice* choice, the verb is inflected with different thematic suffixes. I will not discuss them here since I did not encode this information into the analyzer. For a detailed analysis on this subject, see Öztürk (2021) and Taylan and Öztürk (2014). Unlike other thematic suffixes, *-e(r)* is very predictable and always realized with *i-* valency-related vowel in derived unaccusatives unless it is not overwritten.

3.5.1.2.8 Subjunctive marker

Subjunctive (SUBJ) suffix *-a* can be used to express future tense, subjunctive mood as well as epistemic and deontic modalities preceding person markers expressing present tense and also auxiliaries *-(e)re* and *-ertu*. See Öztürk and Pöchtrager (2011) for more information and examples on the use of subjunctive marker *-a*.

Morphotactic constraint 16: The subjunctive marker *-a* only precedes present tense denoting person suffixes.

3.5.1.2.9 Conditional marker

There are two conditionals (COND) in Pazar Laz as *real* and *unreal* conditionals.

Unreal conditional is formed with the *-k'o* suffix occurring between the past tense set of person markings and plural markers. Interestingly, when being used with the 3rd person past tense and plural denoting suffix *-es*, there is an additional past tense marking *-e* before *-k'o* as seen in 89.

(89) *Ordo-şe* *m-i-lv-ap-u-t'-e-k'-es* *hus*
early-ABL D-P.1-VAL-come-CAUS-TS-IMPF-PST-COND-PRS.3PL now
noğa *mok'oveleri* *v-ort'-a-Ø-t-ert'u*
market.NOM crossed S-A.1-COP-SUBJ-PRS.1-PL-AUX.PAST
'If we had gone earlier, we would have crossed the market by now.'

Source: Bucak'lişi et al. (2007)

Morphotactic constraint 17: The conditional suffix *-k'o* only follows the past tense set of person suffixes and precedes plural markers. In case the person suffix denotes 3rd person and plurality, *k'o* is followed by an extra *-e* past tense suffix. *-e* is only possible before the unreal conditional *-k'o*.

The suffix *-na*, which comes after tense, person, and number suffixes, expresses the real condition. It can occur with present and past tense set of person suffixes as well as denote future tense with the subjunctive marker *-a*.

3.5.1.2.10 Auxiliaries

There are two auxiliaries (AUX) attached to the verbal complex following plural markers at position 11 as *-(e)re* and *-ert'u*.

3.5.1.2.11 Other affixes and clitics/particles for fully inflected verbs

In Öztürk and Pöchtrager (2011), there are some particles that attach to the fully inflected verbal complex at the end such as question clitic *-i* for polar and echo questions and *-do* for rhetorical questions. Additionally, the additive *-ti* can also attach to the verbal complex for conjunction. These particles are found to directly attach to or to be separated from the verbal complex with a hyphen.

Finally, in case of adverbial clauses, the embedded fully inflected verb can take suffixes such as *-(s)is*, *-şa* (also ALL), *-şe(n)* (also ABL) to denote 'when', 'while'/'until' and 'before' respectively.

Subordinators *ya do* and *ma do* are defined to express the meaning of 'so that' when following the fully inflected verb in the embedded sentence. The choice between these two forms depends on the person information of the subject in the matrix clause. If the subject is first person, then *ma do* is selected. These *ya* and *ma* subordinators together with *şo* on their own are also used to indicate direct speech. In Bucak'lişi et al. (2007), they can be found to attach with a hyphen to the verb similar to question clitics.

The subordinator *na* is used in complement clauses, relative clauses and purpose adverbial clauses. In these constructions, *na* precedes the embedded verb and can attach to the verb with a hyphen acting as a prefix. Note that when *na* attaches as a suffix to the verb, it forms and denotes real conditional.

Finally the participle (PART) *-eri* is used very productively to produce manner adverbials. It attaches directly to the verb root.

3.5.2 Summary of morphotactic rules and constraints

This section presents the collection of morphotactic rules and constraints introduced throughout this chapter.

3.5.2.1 Nouns

The general order of inflectional affixes for noun stems is given in 90.

(90) noun stem + plurality + possessive + postposition + case + additive

Morphotactic constraint 1: Only certain lexemes ending with *a* can be marked with -*lepe* plural marking[48].

3.5.2.2 Verbs

The positions for verbal inflectional affixes are given in an order in 91 in relation to the root. The parentheses introduce the additional affixes and clitics that mark the verbal complex defined in Section 3.5.1.2.11.

(91) (subordinator |) affirmative preverb | spatial preverb | person prefix | valency-related vowel | ROOT | augmentative | causative intr. | causative tran. (| participle) | causative perf. | thematic suffix | imperfective | subjunctive | person suffix | conditional | plurality | auxiliary (clitics; additive, question | adverbial | subordinator)

[48] The dictionary by Bucak’lişi et al. (2007) provides this information in the word definitions. They are extracted and labeled based on the existence of the -*lepe* suffix in the word entries.

Morphotactic constraint 2: If preverbal (-2) S-A person markers are realized on the verbal complex, they need to reflect the same person value as the postverbal (8) S-A person markers.

Morphotactic constraint 3: Disjoint person markings cannot share the same person value except 3rd person.

Morphotactic constraint 4: If the position 8 is marked with 1st person, the position -2 is never empty. In this case, the unmarked P-D.3 prefix is disallowed before the S-A.1 suffix.

Morphotactic constraint 5: If the person suffix denotes 1st or 2nd person, the plural marking *-t* should follow the person suffix. Plurality with 3rd person only allows fused forms and prevents marking of the singular counterparts given in Table 10 before plural 3rd person suffixes.

Morphotactic constraint 6: In case of applicatives such as *Appl: Recipient, Appl: Possessor, Appl: Benefactor, Appl: Experiential perfect*, if the *i-* vowel fills the position -1, it is always preceded by either the P-D.1 person prefix (*m-*) or a P-D.2 person prefix (*g, k, k'*), and if *u-* is used, the person prefix can be unmarked reflecting the value of D-P.3 or can be in joint agreement with the person prefix at position 8, and overtly marked only if the person prefix is 1st person[49].

Morphotactic constraint 7: Under *Appl: Experiential perfect* construction, following the vowel *i-/u-* the verb complex needs to be marked with the causative suffix *-ap* and the thematic suffix *-u(r).*

Morphotactic constraint 8: The verb stem is marked with the valency related vowel *a-* under the constructions as *Appl: Location*, *Appl: Abilitative, Appl: Deagentive* and *Appl: Psych*. While *Appl: Abilitative*, *Appl: Deagentive* and *Appl: Psych* also require the thematic suffix *-e(r)* to express the backgrounded agentive subject, *Appl: Location* does not need any thematic suffix to be able to surface.

[49] With these instructions, the analyzer will give 3 different outputs when the person prefix is unmarked preceding the *u-* vowel. The hierarchical structure between person prefixes drawn in 45 does not differentiate between unrealized D-P.3, S-A.2 and S-A.3 either. Theoretically, all these three cases are possible with the *u-* vowel.

Morphotactic constraint 9: When causativized and marked with *o-*, the verb is required to be inflected with causative suffixes as *-in* and *-ap*. At least one of them should be present for the verb stem to be able to express the causative.

Morphotactic constraint 10: Under pronominal cliticization, the verb is marked with *i-* and the person markers should be in joint agreement.

Morphotactic constraint 11: Realization of the reciprocal *i-* is dependent on the spatial preverb *ok'o-*.

Morphotactic constraint 12: Under AIC operations, The verb is marked with the pronominal clitic *i-* and the thematic suffix *-e(r)*. In case of unergatives and unaccusatives, it is also marked with the causative marker *-in*. Since there isn't a list of unergatives and unaccusatives available, the *-in* marks every verb stem optionally.

Morphotactic constraint 13: The AIC construction denoting impersonal passive with an applicative benefactor argument requires the verb to be marked with the valency-related vowel *a-* and the thematic suffix *-e(r)*.

Morphotactic constraint 14: When higher applicatives are combined with causatives, the verbal complex is marked with applicative valency-related vowels while showing causative morphology (e.g., causatives suffixes *-in* and *-ap*) in addition to applicative morphology.

Morphotactic constraint 15: The subjunctive marker *-a* only precedes present tense denoting person suffixes.

Morphotactic constraint 16: The conditional suffix *-k'o* only follows the past tense set of person suffixes and precedes plural markers. In case the person suffix denotes 3rd person and plurality, *k'o* is followed by an extra *-e* past tense suffix. *-e* is only possible before the unreal conditional *-k'o*.

3.6 Morphophonology

3.6.1 Nouns

The final *i* sound of noun stems becomes *e* when the stem is inflected with the plural marker *-pe*. See Section 3.5.1.1 for examples.

3.6.2 Verbs

3.6.2.1 Verb stem

Verb stems ending with *v* alternates between Ø and *v*. *v* is deleted when -*u* (3rd person singular past tense marker) attaches to the verb stem. See the verbal inflection in 92 and 93.

(92) *o-v-i-ç'v-i*
PV-S-A.1-VAL-burn-PST.1SG
'I burned.'

Source: adapted from Bucak'lişi et al. (2007)

(93) *dele-m-i-ç'-u*
PV-D-A.1-VAL-burn-PST.3SG
'It burned me.'

Source: adapted from Bucak'lişi et al. (2007)

3.6.2.2 Preverbs

The preverbs show a great amount of morphophonological alternations in their final sounds 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 when followed by vowels *a*, *o*, *i* and *u*. Even though they may end with the same vowel, the alternations can be different when being followed by the same sound. For example, the final *o* sound in *exo*- turns into *v* when it attaches to a verbal complex starting with *a* sound but not the one in *moyo*-. See Öztürk and Pöchtrager (2011) for a detailed examination on these changes.

3.6.2.3 Person prefixes

Person prefixes on the verbal complex usually alternate based on the following consonant's laryngeal property as voiced, voiceless and ejective.

The *v*-set shows alternations of *v*-, *p*-, *p'*-, *b*-, *m*- and *f*- which are not always related to phonology. The *p*-, *p'*- and *b*- alternation can be explained with laryngeal properties of the following consonant. The *v*- form is used if the following sound is a vowel. The *f-* form is observed with the verb stem *xt'*, a suppletive form of the verb *olva* 'move' (Öztürk & Pöchtrager, 2011). While *x* is deleted, the person prefix takes the form *f-*.

Laz also exhibits a sound change in verb stems starting with *n* sound when preceded by ejective *p'-*, the person prefix for 1.SG. The two consonants are combined and becomes *m*. This specific sound change is only defined for the verb stem *nç'ar* 'write' in Öztürk and Pöchtrager (2011).

The *g*-set shows alternations of *k-*, *k'-* and *g-*, purely depending on the laryngeal properties of the following consonant.

3.6.2.4 Thematic suffixes *-e(r)* and *-u(r)*

The final [r] sound in thematic suffixes *-e(r)* and *-u(r)* gets deleted before singular 3rd person marking denoting present tense *-n*. The existence of *-n* also depends on these thematic suffixes. This means that the present form of singular 3rd person suffix also alternates between *-s* and *-n*. See and compare the examples in 94 and in 95.

(94) *K'at'u-k toma go-x-um-s*
cat-ERG fur.NOM PV-fall-TS-PRS.3SG
'Cat sheds fur.'

Source: Bucak'lişi et al. (2007)

(95) *But'k'ape go-y-xv-e-n*
leaves.NOM PV-VAL:AIC-fall-TS-PRS.3SG
'The leaves are falling.'

Source: Bucak'lişi et al. (2007)

3.6.2.5 3rd person plural suffix denoting present tense *-(a)n*

Plural 3rd person suffix *-(a)n* loses the initial *a* following the subjunctive *-a*.

3.6.2.6 Conditional suffix *-k'(o)*

The conditional suffix *-k'(o)* loses the final *o* when followed by vowels. See the sentence in 89 for the deletion of *o* before the person marking expressing 3rd person, past tense and plurality represented as *-es*.

3.6.2.7 Auxiliary *-(e)re*

The auxiliary *-(e)re* loses the initial *e* when preceded by the subjunctive *-a* seen in the example in 96.

(96) *me-f-t'-a-Ø-re*
PV-S-A.1-come-SUBJ-PRS.1SG-AUX
'I will come.'

Source: Bucak'lişi et al. (2007)

3.6.2.8 Adverbial suffix *-(s)is*

The initial *s* sound in the adverbial suffix *-(s)is* realizes only before a vowel as seen in the example in 97 compared to the example in 98.

(97) *Oxori-şe v-i-d-i-sis*
house-ABL S-A.1-VAL-go-PST.1SG-when
'When I went home'

Source: Bucak'lişi et al. (2007)

(98) *Şk'u oxori-şe v-i-d-i-t-is*
we.NOM house-ABL S-A.1-VAL-go-PST.1-PL-when
'When we went home'

Source: Bucak'lişi et al. (2007)

3.7 Conclusion

In this section, I extensively discussed the grammar of Pazar Laz that are relevant to developing a morphological analyzer. I started examining phonetics and phonology of Pazar Laz since Laz does not have a standard orthography which heavily depends on how speakers perceive and produce sounds in the language. This has led me to develop some strategies in order to handle dialectical variations in the language data such as creating a common lexicon for Laz rather than only focusing on words in Pazar Laz dialect. Later, I discussed morphotactics as well as several morphosyntactic properties related to nominal and verbal inflection mainly. I introduced the verbal complex that exhibits both concatenative and non-concatenative morphology. The summary of verbal morphotactics is given in Section

3.5.2. This is basically a compilation of rules and constraints that I specifically focused on when developing the analyzer. The detailed examination of these can also be found in Section 3.5.1.2. Finally, morphophonology is discussed to define rules for producing the surface forms from abstract lexical forms.

CHAPTER 4

METHODOLOGY AND FORMALISM

4.1 Introduction

In this section I will give details how I implemented the morphological analyzer for Laz by using finite-state transducers and two-level morphology with Xerox lexc and twolc formalism.

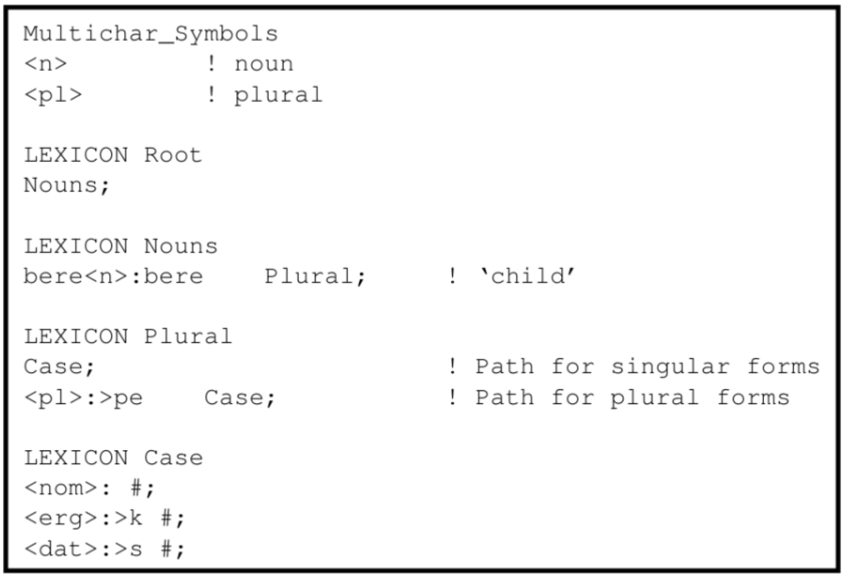
Xerox lexc and twolc formalism are used to create lexicon files and a two-level grammar file in which morphotactic and morphophonemic rules are written Beesley and Karttunen (2003). Open-source Helsinki Finite-State Toolkit (HFST; Linden et al. (2011)) which is popular in this field of research is used to compile these rules into finite-state transducers which can be used as both an analyzer and generator.

I will talk about the syntax of these formalism here but there is also an extensive amount of research available to refer to in the literature for details (see (Beesley & Karttunen, 2003)). I will also discuss and draw attention to a specific technique called *flag diacritics* for their importance in dealing with non-concatenative morphology and overgeneration in Section 4.4.

4.2 Lexc file for morphotactic rules

The *lexc* formalism is used to define a source lexicon for two-level rules in order to compose a final transducer that has all possible word forms in the defined language along with their lexical form (Beesley & Karttunen, 2003).

The *lexc* file consists of two parts in general. See a sample of *lexc* file in Figure 5. The first part is for defining *multichar symbols* that are basically labels or *tags* attached to word stems and affixes usually to mark morphosytactic information. The second part consist of classes of morphemes defined below different *LEXICONs* and connected with *continuation classes* followed by a semi-colon. Two sides separated with a colon are morphotactic and lexical sides.



```
Multichar_Symbols
<n>         ! noun
<pl>        ! plural

LEXICON Root
Nouns;

LEXICON Nouns
bere<n>:bere    Plural;     ! ‘child’

LEXICON Plural
Case;                       ! Path for singular forms
<pl>:>pe    Case;           ! Path for plural forms

LEXICON Case
<nom>: #;
<erg>:>k #;
<dat>:>s #;
```

Figure 5. A fragment of *lexc* file

The lexc file includes verbs and nouns as two main word classes that take inflectional morphemes. Morphotactics regarding these classes are defined according to the discussion in Chapter 3. Also refer to Section 3.5.2 for the summarized version of this discussion. I repeated the order and positions of affixes for noun and verb inflection respectively in 99 and in 100 from Section 3.5.2.

(99) noun stem + plurality + possessive + postposition + case + additive

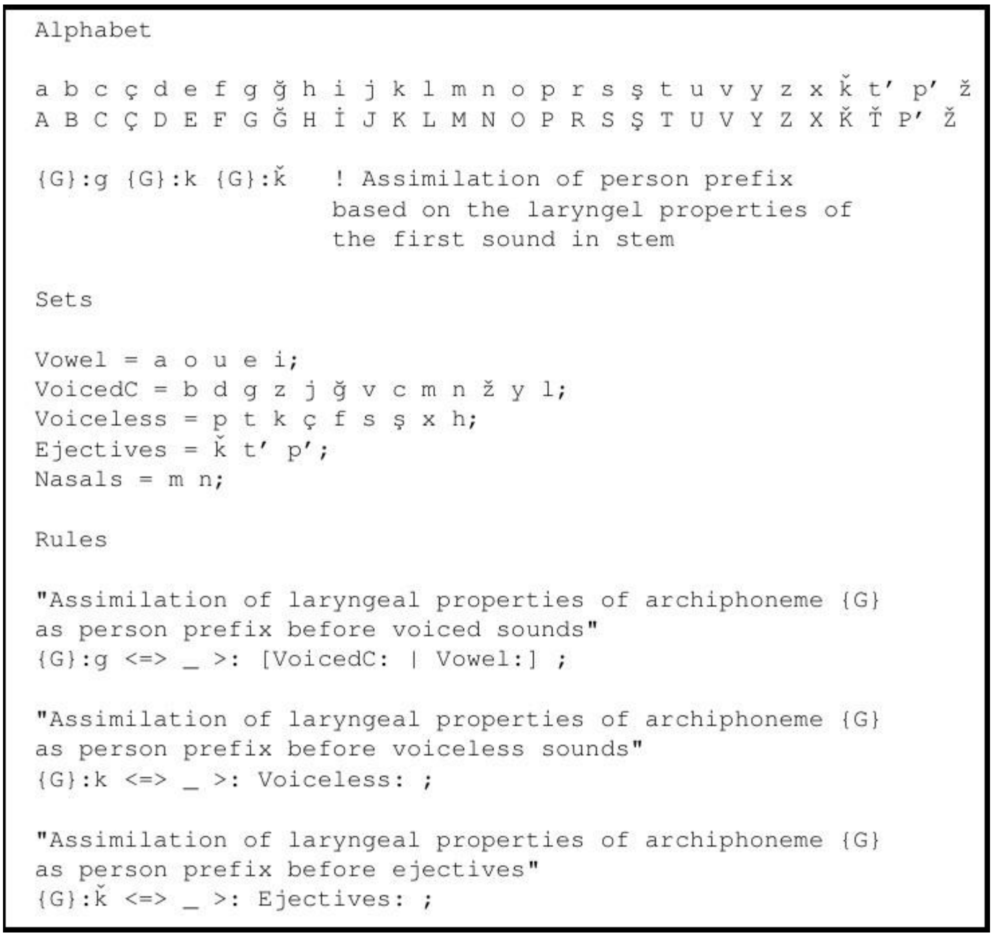
(100) (subordinator |) affirmative preverb | spatial preverb | person prefix | valency-related vowel | ROOT | augmentative | causative intr. | causative tran. (| participle) | causative perf. | thematic suffix | imperfective | subjunctive | person suffix | conditional | plurality | auxiliary (| additive, question, adverbial, subordinator)

The orderings in 99 and 100 basically show the continuation classes inside the *lexc* file. For the morphotactic tags I have mostly followed the description in Öztürk and Pöchtrager (2011) as I discussed in Chapter 3.

In addition, the *lexc* file has other substantives such as adjectives and adverbs as well as pronouns as personal pronouns, possessive pronouns, demonstratives, reflexives, interrogative pronouns, indefinite pronouns. Interjections, quantifiers, adpositions, conjunctions, subordinators, negations and numerals are also included from Öztürk and Pöchtrager (2011) and Bucak'lişi et al. (2007). The size of the lexicon will be discussed in Section 5.2 in Chapter 5.

4.3 Twolc file for morphophonological rules

The *twolc* file has three main parts as *alphabet*, *sets* and *rules*. See a small part of *twolc* file in Figure 6. Alphabet is used to define all letters in the language together with *archiphonemes* on which phonological rules apply. Sets are to prevent accumulation of letters inside the rules and usually defined for natural classes of sounds such as vowels, voiced and voiceless consonants.



```
Alphabet

a b c ç d e f g ğ h i j k l m n o p r s ş t u v y z x ǩ t' p' ž
A B C Ç D E F G Ğ H İ J K L M N O P R S Ş T U V Y Z X Ǩ Ť P' Ž

{G}:g {G}:k {G}:ǩ   ! Assimilation of person prefix
                    based on the laryngel properties of
                    the first sound in stem

Sets

Vowel = a o u e i;
VoicedC = b d g z j ğ v c m n ž y l;
Voiceless = p t k ç f s ş x h;
Ejectives = ǩ t' p';
Nasals = m n;

Rules

"Assimilation of laryngeal properties of archiphoneme {G}
as person prefix before voiced sounds"
{G}:g <=> _ >: [VoicedC: | Vowel:] ;

"Assimilation of laryngeal properties of archiphoneme {G}
as person prefix before voiceless sounds"
{G}:k <=> _ >: Voiceless: ;

"Assimilation of laryngeal properties of archiphoneme {G}
as person prefix before ejectives"
{G}:ǩ <=> _ >: Ejectives: ;
```

Figure 6. A fragment of *twolc* file

Each rule defines the context for an archiphoneme to be able to surface as a specific sound. The rules for D-P.2 person prefix alternations are introduced in Figure 6. The rules inside the *twolc* file are based on the morphophonological alternations discussed in Section 3.6 under Chapter 3.

4.4 Flag diacritics

Although FSTs can turn into very robust morphological analyzers, they tend to overgenerate. In order to prevent overgeneration, morphotactic rules should be defined clearly. However, not all morphotactic rules can be written in a concatenative manner in continuation classes.

Figures 1 and 2 introduced in Section 2.3.1 and 2.3.2 under Chapter 2 do not explain how to handle long-distance dependencies on the word-level, and not all languages have concatenative morphotactics or fully concatenative morphotactics as in case of Laz. Laz exhibits non-concatenative, templatic morphology (Atlamaz, 2013). It inflects the verb complex with suffixes and prefixes which are sometimes used together forming long-distance dependencies on the word-level.

In order to write constraints over long-distance dependencies, Xerox provided a set of feature-setting and feature-unification operations, namely *flag diacritics* that would eliminate paths that are not desirable during compilation (Beesley & Karttunen, 2003). Instead of creating new rules in the FST, which would only make the transducer bigger, applying flag diacritics can be very efficient, by preventing the transducer to increase in size. Additionally, not only long-distance dependencies but also idiosyncrasies of roots such as selecting a specific type of affix which does not depend on morphophonology can be dealt with flag diacritics. Flag diacritics are set in a format provided in 101 and only exist in the lexc file but not twolc file. They label paths both at the lexical side and the surface side. There are several flag types as shown in 102 that have different functions.

(101) @FLAGTYPE.FEATURE.VALUE@

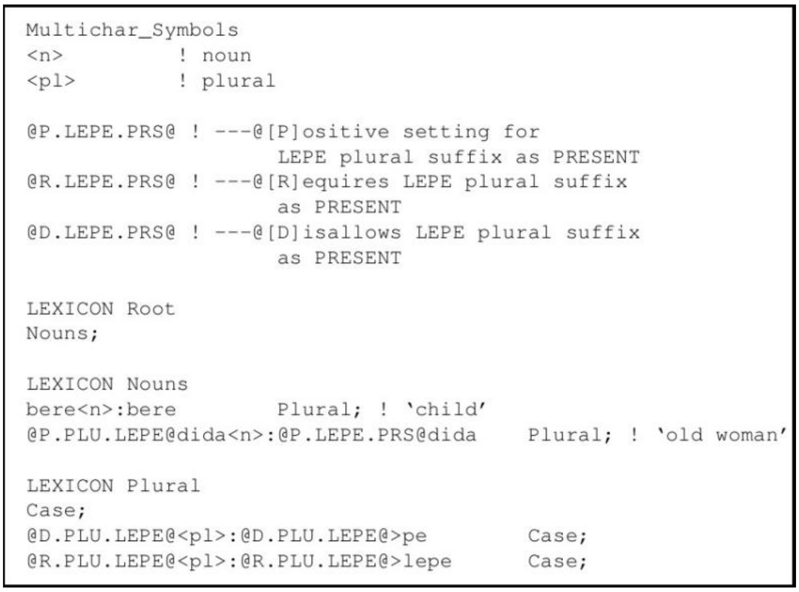
(102) Flag types and their usage

U : [U]nification Unifies paths that have the same FEATURE and VALUE
P : [P]ositive Sets paths as [P]ositive for specified FEATURE and VALUE
R : [R]equire [R]equires paths set as [P]ositive for specified FEATURE and VALUE
D : [D]isallow [D]isallows paths marked [P]ositive with specified

	FEATURE and VALUE
N : [N]egative	Sets paths as [N]egative for specified FEATURE and VALUE FEATURE and VALUE

Source: Beesley and Karttunen (2003)

An example case for the use of flag diacritics is given below in Figure 7. The [P]ositive feature operation for the plural marker *-lepe* is defined for the noun stem *dida*, but not for *bere*. In addition, *-lepe* is labeled with [R]equire feature operation for the feature LEPE and the value as PRS while *-pe* is labeled with Disallow operation for the feature LEPE and the value as PRS. Therefore, the *-pe* plural suffix will not accept paths marked with @P.LEPE.PRS@ while the *-lepe* plural suffix will only accept paths marked with @P.LEPE.PRS@. Based on this sample lexc file, the paths allowed are provided in 103 and those that are ignored are given in 104.



```
Multichar_Symbols
<n>          ! noun
<pl>         ! plural

@P.LEPE.PRS@ ! ---@[P]ositive setting for
                    LEPE plural suffix as PRESENT
@R.LEPE.PRS@ ! ---@[R]equires LEPE plural suffix
                    as PRESENT
@D.LEPE.PRS@ ! ---@[D]isallows LEPE plural suffix
                    as PRESENT

LEXICON Root
Nouns;

LEXICON Nouns
bere<n>:bere        Plural; ! 'child'
@P.PLU.LEPE@dida<n>:@P.LEPE.PRS@dida    Plural; ! 'old woman'

LEXICON Plural
Case;
@D.PLU.LEPE@<pl>:@D.PLU.LEPE@>pe        Case;
@R.PLU.LEPE@<pl>:@R.PLU.LEPE@>lepe      Case;
```

Figure 7. The use of flag diacritics in lexc file

(103) bere@D.LEPE.PRS@>pe

@P.LEPE.PRS@dida@R.LEPE.PRS@>lepe

(104) bere@R.LEPE.PRS@>lepe

@P.LEPE.PRS@dida@D.LEPE.PRS@>pe

This solves the 1st morphotactic constraint defined in Chapter 3.

Morphotactic constraint 1: Only certain lexemes ending with *a* can be marked with -*lepe* plural marking

Laz verb complex has required substantial use of *flag diacritics* 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[1].

How to handle joint and disjoint person markings using flag diacritics can be seen in Figure 8. Since affixation starts with person prefixes[2], they are set with [P]ositive flag for features S, marking sole arguments and D-P, marking dative-marked or patient argument, and for values 1st, 2nd and 3rd person. The constraints take place by using [D]isallowing flags on person suffixes. For example, the person prefix denoting past tense singular 1st person realizing as -*i* does not accept paths marked with 2nd and 3rd person subject markings. Note that this specific suffix also disallows 1st person D-P person prefix due to the 3rd morphotactic constraint defined in Chapter 3 for disjoint person markings.

Morphotactic constraint 3: Disjoint person markings cannot share the same person value except 3rd person.

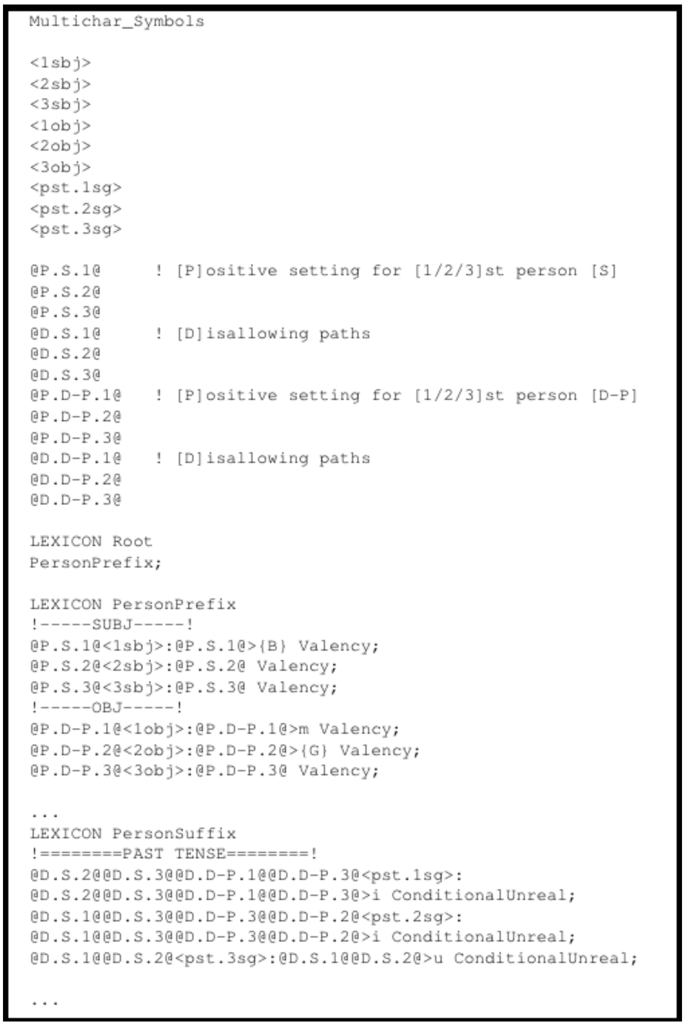
The 4th morphotactic constraint is also handled by adding @D.D-P.3@ for the same person 1st person suffix.

Morphotactic constraint 4: If the position 8 is marked with 1st person, the position -2 is never empty. In this case, the unmarked P-D.3 prefix is disallowed before the S-A.1 suffix.

Flag diacritics are useful for overwriting of valency-related vowels when two separate operations/constructions apply 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. Laz allows overwriting causative *o*- to be overwritten by applicative *i*- while keeping post verbal causative markers -*in* or -*ap*.

[1] These constructions and specific morphotactic constraints are extensively discussed in Chapter 3.

[2] Note that this is a simplified version of the original *lexc* file. The continuation classes and their order are given in 100 in Section 4.2. Also person prefixes are only given for the past tense paradigm.



```
Multichar_Symbols

<1sbj>
<2sbj>
<3sbj>
<1obj>
<2obj>
<3obj>
<pst.1sg>
<pst.2sg>
<pst.3sg>

@P.S.1@      ! [P]ositive setting for [1/2/3]st person [S]
@P.S.2@
@P.S.3@
@D.S.1@      ! [D]isallowing paths
@D.S.2@
@D.S.3@
@P.D-P.1@   ! [P]ositive setting for [1/2/3]st person [D-P]
@P.D-P.2@
@P.D-P.3@
@D.D-P.1@   ! [D]isallowing paths
@D.D-P.2@
@D.D-P.3@

LEXICON Root
PersonPrefix;

LEXICON PersonPrefix
!-----SUBJ-----!
@P.S.1@<1sbj>:@P.S.1@>{B} Valency;
@P.S.2@<2sbj>:@P.S.2@ Valency;
@P.S.3@<3sbj>:@P.S.3@ Valency;
!-----OBJ-----!
@P.D-P.1@<1obj>:@P.D-P.1@>m Valency;
@P.D-P.2@<2obj>:@P.D-P.2@>{G} Valency;
@P.D-P.3@<3obj>:@P.D-P.3@ Valency;

...
LEXICON PersonSuffix
!========PAST TENSE========!
@D.S.2@@D.S.3@@D.D-P.1@@D.D-P.3@<pst.1sg>:
@D.S.2@@D.S.3@@D.D-P.1@@D.D-P.3@>i ConditionalUnreal;
@D.S.1@@D.S.3@@D.D-P.3@@D.D-P.2@<pst.2sg>:
@D.S.1@@D.S.3@@D.D-P.3@@D.D-P.2@>i ConditionalUnreal;
@D.S.1@@D.S.2@<pst.3sg>:@D.S.1@@D.S.2@>u ConditionalUnreal;

...
```

Figure 8. The use of flag diacritics for joint and disjoint person markings (simplified)

These cases require the overwriting constructions to have additional flags. For applicatives to get the causative markers, the applicative *i-* (as well as causative *o-*) need to have the flag @P.CAUS.PRS@ which will let them through paths defined with @R.CAUS.PRS@ which requires the causative feature to be present. This labelling with flag diacritics handles the 14th morphotactic constraint.

Morphotactic constraint 14: When higher applicatives are combined with causatives, the verbal complex is marked with applicative valency-related vowels while showing causative morphology (e.g. causatives suffixes -*in* and -*ap*) in addition to applicative morphology.

All other morphotactic constraints defined in Section 3.5.2 under Chapter 3 are the specific rules that are solved with flag diacritics as well.

4.5 Compilation into transducers

After adding all the morphotactic and morphophonological rules into the *lexc* and *twolc* files, the next step is to compile these files into one transducer that will work as a morphological analyzer. HFST compilers and command line tools are used for this study.

hfst-lexc and *hfst-twolc* compilers take the *lexc* and *twolc* files separately and output two different FSTs. An important part is to combine these two files in order to get surface forms of the lexical forms defined in the lexicon. *hfst-compose-intersect* is the compiler for this purpose. The output is the generator. *hfst-invert* is used to turn this generator into an analyzer.

Lastly, the most important tool especially for the end users of this analyzer is *hfst-lookup*. This tool is used to analyze a given word form with a morphological analyzer compiled with *hfst* tools. Either a single token or a file containing each token in a separate line is accepted. The tool reads a given input line-by-line.

4.6 Conclusion

In this chapter, I introduced the methodology and formalism in which morphotactics and morphophonological rules are written and compiled into transducers (as generator and analyzer). I also discussed the importance of using flag diacritics for their advantage over non-concatenative morphology and explained how to use flag diacritics to handle long-distance dependencies on the word-level such as joint and disjoint person marking.

CHAPTER 5

EXPERIMENTS AND RESULTS

5.1 Introduction

In this chapter, I will give the results of the morphological analyzer after I discuss the corpus collection and lexicon preparation for the source lexicon of the FST. The corpus was needed to evaluate the analyzer. I also prepared a lexicon containing word stems to include in the lexc file in order to increase the coverage of the analyzer.

I will finally talk about some of the challenges I have experienced during the development of the analyzer. Since before this study, there were no computational language tools or resources, I had to prepare everything from scratch, which includes the lexicon building. This was the hardest and the most time-consuming part of the study which would still be improved.

5.2 Lexicon

The lexicon composed for this study mostly comes from *Büyük Lazca Sözlük* (Didi Lazuri Nenapuna). It is the most extensive dictionary available for Laz prepared by İsmail Bucak'lişi, Hasan Uzunhasanoğlu 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[1] and if available, dialect-specific forms as seen in (5.2). (5.2) shows that the lexical entry *doinu* 'to give birth' belongs to *Atn.* and *Viw.* dialects, and it takes the form of *dorinu* in *Gyl., Ark.* and *Xop.* dialects, and *dok'unapa* in *Sap.* dialect.

(105) doinu Atn., Viw., dorinu Gyl., Ark., Xop., dok'unapa Sap.

[1] The following dialect codes were found in the dictionary: *'Yul' (Eastern dialects), 'Gyl' (Western dialects), 'Viw' (Fındıklı), 'Xop' (Hopa), 'Ark' (Arhavi), 'Çxl' (Borçka-İçkale), 'Atn' (Atina/Pazar), 'Fur' (Çamlıhemşin), 'Arş' (Ardeşen), 'Sap' (Sapanca).*

I prepared 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, I 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 & Karttunen, 2003) (see the discussion in 3.2.3, 3.3 and 3.4 in Chapter 3).

The extraction of verbs from the dictionary was based on the Turkish translation of the words. Those which end with the *-mAk* infinitive marker in Turkish indicate that they have a verbal stem. The reason why I did not look at Laz verb endings such as *-u* for extractions is the fact that Turkish is more regular in verb endings and more error-proof. Although it was easy to find verbs this way, 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.

Collecting substantives from the dictionary and carefully separating them into nouns, adverbs, and adjectives were other difficult tasks for this study. Arranging words into their corresponding word classes was done semi-automatically and there were words that should be put in more than one category such as noun and adjective or adjective and adverb (determined only by sentential position) or noun and adverb (see Öztürk and Pöchtrager (2011) for substantives). Categorizing other syntactic elements like interjections, conjunctions and pronominals were done manually.

One of my thesis committee members, Dear Assoc. Prof. Ümit Atlamaz suggested me to add Turkish words in order to cover loanwords found in corpora in a substantial amount. Instead of directly adding Turkish words as potential loanwords bearing the phonological properties of Laz such as nouns ending in [i], [ı] changing to [i] or voiceless stops becoming ejectives, which would increase the size of the lexicon unnecessarily and possibly lead to having incorrect forms in the lexicon, I decided to

Table 19. Number of Lexicon Entries by Part of Speech / Lexical Category.

Category	**Number of Stems**
Noun	17,589
Verb	2442
Adjective	2422
Adverb	1198
Pronoun	246
Numeral	164
Interjection	144
Postposition	78
Quantifier	56
Conjunction	31
Negation	5
Preposition	3
Total	24,378

use the corpora as another source for the lexicon building. By using such phonological properties of Laz I extracted Turkish loanwords from the corpora. This increased the number of noun stems by 1915 including some original Laz word stems as well. Some Turkish loanwords are given in 106 which are not found in the dictionary but the corpora only.

(106) *mangalciluği* (Turkish: ‘mangalcılık’, English: ‘barbecue’)
xirsuzluği (Turkish: ‘hırsızlık’, English: ‘theft’)
k’omşiluği (Turkish: ‘komşuluk’, English: ‘neighborliness’)
balk’oni (Turkish: ‘balkon’, English: ‘balcony’)
k’iyamet’i (Turkish: ‘kıyamet’, English: ‘apocalypse’)
ayak’abi (Turkish: ‘ayakkabı’, English: ‘shoe’)
çarçafi (Turkish: ‘çarşaf’, English: ‘sheet’)

5.3 Corpus

I have collected different type of written texts for the Laz corpus. However, due to differences in terms of dialects these texts are divided into their corresponding dialects for this preliminary study. Unfortunately, Pazar Laz has almost no written text known in the literature. The only resource was the courtesy of my informant, İsmail Bucak'lişi, an 800 page document consisting of 110,417 tokens collected by İsmail Bucak'lişi, a native speaker of Pazar Laz, by himself which contains daily conversations and stories shared in his immediate circle. It should be noted that the original document also contains Turkish words and sentences given as translations throughout the document. However, after using a Turkish morphological analyzer over the corpus, I eliminated 26,864 potential Turkish words. This makes the total token count 83,553. I will call this corpus as PLC (Pazar Laz Corpus).

The other corpus is collected from texts written in Fındıklı dialect which is an eastern dialect. Therefore, it is called FLC (Fındıklı Laz Corpus) which contains poems, short stories written by Nurdoğan Demir Abaşişi and the translation of 'White Fang' (Jack London) by Osman Şafak Buyuklişi. The token count for FLC is 109,355.

Another resource was a very recent one developed under Universal Dependencies (Nivre et al., 2017). It is a small treebank consisting over 576 sentences and 2,306 tokens for Pazar Laz collected from grammar books, theses and research articles (Türk, Bayar, Özercan, Öztürk, & Özateş, 2020). However, this could not be used for training a machine learner due to its size and not all sentences in the treebank are glossed for morphological features. Additionally, I noticed that it contains orthographic inconsistencies such as 'ş' written either as 'š' and 'sh' as well as spelling mistakes such as *berepe* 'children' written as **bereepe* and some Georgian words and sentences. After cleaning these, the number of tokens decreased to 1872.

5.4 Results

The morphological analyzer has been evaluated by calculating the naïve coverage and doing error analysis on randomly selected 100 tokens from PLC. See an example output of the analyzer in 108 for the sentence in 107.

(107) *Ar xordza-k bere-pe muşi kononciru*
one woman-ERG child-PL.NOM her PRV-PRV-VAL:CAUS-sleep-PST.3SG
do oxori ek'isvaramt'u
and house.NOM PRV-VAL-tidy.up-TS-IMPF-PST.3SG
'A woman has put her children to sleep and was tidying up the house.'

Source: Öztürk and Pöchtrager (2011)

(108) Output of the analyzer

xordzak	xordza<N><ERG>
berepe	bere<N><PL><NOM>
muşi	muşi<PRO><3.SG><GEN>
kononciru	<PRV.AFFR><PRV.SPTL><P-D.3><VAL>ncir<VERB><PST.3SG>
do	do<conj>
oxori	oxori<n><nom>
ek'isvaramt'u	<˙ PRV.SPTL><P-D.3><VAL>svar<VERB><TS><IMPF><PST.3SG>

5.4.1 Naïve Coverage

Table 20. Naïve Coverage of The Analyzer

Corpus	Tokens	Coverage	Unique Tokens	Coverage
PLC	83,553	%89.9	28,757	%78.2
Laz Treebank	1872	%95.9	632	%92.1
FLC	109,355	%84.6	21,134	%74.3

The coverage is measured by calculating the number of the tokens that receive at least one morphological analysis by the analyzer. It should also be noted that the tokens may have other analysis that is correct but not provided by the analyzer even though they get at least one analysis. The final morphological analyzer has the following coverage percentages over three different corpora developed for Pazar Laz (i.e., PLC and Laz Treebank) and Fındıklı dialect (i.e., FLC). Coverage is calculated for all tokens as well as unique tokens in the corpora.

5.4.2 Error Analysis

I have looked at the tokens that are not covered by the morphological analyzer. Randomly selected 100 tokens from PLC have been examined and separated according to their error type seen in Table 21. It should also be noted that some of them may go into more than one category.

I have chosen PLC primarily for error analysis because the analyzer is based on Pazar Laz and it contains real utterances in Pazar Laz unlike Laz Treebank which may have artificially produced sentences. Also the analyzer has a high coverage over the treebank and tokens that are not analyzed are due to missing stems which are not found in the dictionary such as *metali* ‘metal’ and *past’a* or *pasta* ‘cake’.

The highest percentage of unrecognized tokens belongs to the category of ‘Missing lemma’. Although I have extended the lexicon by adding more word roots from the dictionary, the analyzer still suffers from this issue. I have found that some of the words are not defined in the dictionary as separate word entries (although they are still used in example Laz sentences in the dictionary). For example, one of the unrecognized words was *şk’la* ‘with’ and the dictionary entry for this word only contains these variations; *k’ala*, *şk’ala*. A similar example is *şk’mi* ‘my’. Dictionary entries are *çkimi* and *şk’imi* although *şk’mi* is used in a sentence in the dictionary.

There are unrecognized tokens such as *şukale* ‘after’ in PLC which is found in the dictionary only as *şuk’ale.* These voiceless stop [k] and ejective stop [k’] alternations are very common in the corpora. Another similar example is *kapula* ‘back’ which is entered into the lexicon as *k’apula* from the dictionary.

Additionally, since the collection of word stems for the lexicon were not fully manual, it is prone to error, and it is possible to miss certain word entries due to the organization of the dictionary. In fact, I have come across several times that some word entries from the dictionary were not in separate lines when I turned it into the TXT format for processing. This was a problem since I considered each line as separate word entries; therefore some words were naturally lost if they were not separated into different individual lines.

‘Loanwords’ are in fact another missing lemma category. Although I added a high number of loanwords from the corpora into the lexicon, since tokens are inflected word forms, some of loanwords were missed due to its word form when searching the corpora for potential word lemmas. For example, the loanword *nazart’aşi* ‘nazar taşı’ in Turkish and ‘evil eye stone’ in English is only observed with the instrumental case *-te* in the corpora, and not in its bare form; therefore, it was overlooked during the automatic collection process.

Missing morphotactic and morphophonemic rules are mostly due to productive derivations and unpredicted phonological and morphophonological changes that are not defined for Pazar Laz, respectively.

Table 21. Error Analysis for Randomly Selected 100 Tokens

Error Type	Frequency	Percentage
Missing lemma	47	45.2%
Loanwords	23	22.1%
Typing errors	16	15.4%
Turkish word	8	7.7%
Missing or erroneous morphotactic rule	5	4.8%
Missing or erroneous morphophonemic rule	5	4.8%
Total	104	

Unrecognized tokens from FLC are also mostly due to missing lemmas. However, different morphemes and morphophonological rules in Fındıklı or sometimes east dialects in general increased the number of unrecognized tokens for this corpus specifically. For example, one of the most noticeable difference was the *b-*

alternation of 1st person prefix before vowels. In Pazar Laz, in this context, the person prefix takes the form of *v-*. See the change in the form of the person prefix in 109 for Pazar and Fındıklı dialects.

(109) *v-i-bgar (Pazar)*
b-i-bgar (Fındıklı)
S-A.1-VAL-cry

'I cry/am crying.'

Source: Bucak'lişi and Kojima (2003)

Other examples are the -*z* form of dative case (-*s* in Pazar Laz) and the -*ş* form of genitive case (-*şi* in Pazar Laz).

5.5 Challenges

There were many challenges to this study some of which I already discussed throughout this thesis. Most of them are due to Laz being an endangered language and having no standard written or spoken form. I decided to collect them under some topics.

5.5.1 Little to no language resources for NLP or CL tasks

The first challenge was to prepare every resource I would need for the analyzer from scratch. The first time I started developing this analyzer in 2019, there was nothing except the dictionary (Bucak'lişi et al., 2007), and grammar books. The lexicon is collected with mostly manual categorization and addition of word roots. The word roots are from the dictionary, grammar books and also the corpora I collected afterwards.

There was no corpus available. Only in 2020, a treebank was prepared for Pazar Laz (Türk et al., 2020) which contains 1872 tokens.

5.5.2 Data pre-processing

Preparing language data ready for processing was another issue. Almost every language resource follows a different orthographical system which is also discussed in Section 3.3 and 3.4 in Chapter 3. Therefore, every book or document were pre-processed separately.

Collecting verb roots from the dictionary also required an extensive amount of effort and pre-processing due to verb entries being in their infinitival forms and being already marked with preverbs. Since the analyzer need the root forms, these verb entries from the dictionary had to go through a process of stemming with additional supervision.

5.5.3 Dialectical variations

Dialectical variations I observed in any data I have for this study were incredible. The fact that Laz has no standard written form allows Laz speakers to be flexible when transcribing their speech. Although the dictionary provides an extensive collection over these variations, it came short to create a common lexicon for Laz. Borrowings of words from other dialects are highly evident in the corpora. Therefore, I decided to create a common lexicon for Laz and not Pazar Laz only.

5.5.4 Turkish loanwords

Loanwords are mostly from Turkish and not all of them appear in the dictionary. This has created problems when calculating the coverage of the analyzer since the unrecognized tokens includes a high number of Turkish words abundantly used in Laz with specific sound alternations and additions such as the addition of [i] sound at the end of nouns and changing voiceless stops (e.g., [k]) into ejectives (e.g., [k’]). Therefore, one of the solutions was to use the corpora as a direct source for adding loanwords into the lexicon.

5.6 Developing a common morphological analyzer for Laz

The morphological analyzer for this study is developed on morphotactics and morphophonological rules mostly specific to the Pazar dialect although the lexicon was prepared to cover all dialects.

What I have learned so far led me to support the idea of creating a common analyzer for Laz rather than separate analyzers for each dialect. There are several reasons behind this. The main one is the fact that not every dialect has its own specific grammar described in detail which would be enough to develop an analyzer. Also, most of the morphotactics are the same although the difference between dialects is most evident in phonological and morphological alternations of morphemes. As I emphasized many times throughout this thesis, there is no clear cut lines between dialects. Borrowings of words and word forms are a common phenomenon for Laz dialects.

One important point to keep in mind when developing such an analyzer would still be being aware of dialectical variations and if necessary, encoding them into the lexicon. If a dialect has a very specific morphotactic and morphophonological rule, I suggest adding this in the documentation of the analyzer with additional comments or labels. When needed later, these differences could be referred to in the future.

For example, I wrote additional rules in the final version of the lexicon and two level grammar file also given in 110 and 111 respectively. Being compared to {B} archiphoneme introduced in Figure 8 for the 1st person for S-A arguments, {B2} realizes as *b*- also before vowels, and not just 'VoicedC' set (voiced consonants).

(110) @P.S.1@<1SBJ>:@P.S.1@>{B2} Valency; ! Fındıklı dialect

(111) "Assimilation of laryngeal properties of archiphoneme {B2} as person prefix before voiced consonants and vowels (Fındıklı dialect)"
{B2}:b <=> _ [VoicedC: | Vowel:] ;

5.7 Conclusion

This thesis has presented the very first implementation of a morphological analyzer for Laz leveraging the power of finite-state networks and two-level morphology. The results are promising to consider that Laz does not have a standard written (or spoken) form and exhibits substantial dialectical variations.

Project files including the *lexc* and *twolc* files have been uploaded on Github[2] and licensed under the Creative Commons BY-NC-SA 3.0.

[2] https://github.com/esraonal/laz-morphological-analyser-fst.git

REFERENCES

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W. & Mohri, M. (2007). OpenFst: A general and efficient weighted finite-state transducer library. In Holub J. & Žďárek J. (Eds.), *Implementation and Application of Automata* (pp. 11–23). doi:10.1007/978-3-540-76336-9_3

Anderson, R. D. (1963). *A grammar of Laz* (Doctoral dissertation). The University of Texas, Austin, TX.

Asillazanci, Z. (2018). *Understanding the problems of the support of an endangered language in typography: Proposal of a typeface that supports Laz language* (Doctoral dissertation). Escola Superior de Arte e Design de Matosinhos.

Atlamaz, Ü. (2013). Cyclic agreement and empty slots in Pazar Laz. In C. Cathcart, S. Kang & C. S. Sandy (Eds.), *Proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society: Special Session on Languages of the Caucasus* (pp. 18–31). Retrieved from https://escholarship.org/uc/item/4s07523p

Beesley, K. R. & Karttunen, L. (2003). *Finite state morphology*. Stanford, CA: CSLI Publications.

Beguš, G. (2021). Segmental phonetics and phonology in Caucasian languages. In M. Polinsky (Ed.), *The Oxford Handbook of Languages of the Caucasus* (pp. 687-728). New York, NY: Oxford University Press. doi:10.1093/oxfordhb/9780190690694.013.18

Bird, S. (2009). Natural language processing and linguistic fieldwork. *Computational Linguistics*, *35*(3), 469–474. doi:10.1162/coli.35.3.469

Blevins, J. P. (2016). *Word and paradigm morphology*. Oxford, United Kingdom: Oxford University Press.

Boeder, W. (2005). The South Caucasian languages. *Lingua, 115*(1-2), 5-89. doi:10.1016/j.lingua.2003.06.002

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. doi:10.1162/tacl_a_00051

Booij, G. (2012). *The grammar of words: An introduction to linguistic morphology* (3rd ed.). Oxford, United Kingdom: Oxford University Press.

Bucak'lişi, İ. & Kojima, G. (2003). *Laz grammar (Lazuri grameri)*. İstanbul, Turkey: Chiviyazilari.

Catford, J. C. (1977). Mountain of tongues: The languages of the Caucasus. *Annual Review of Anthropology*, *6*(1), 283–314. Retrieved from http://www.jstor.org/stable/2949334

Ćavar, M., Ćavar, D. & Cruz, H. (2016). (2016). Endangered Language Documentation: Bootstrapping a Chatino Speech Corpus, Forced Aligner, ASR. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., & Piperidis, S. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4004–4011). European Language Resources Association (ELRA).

Ç’ikobava, A. (1936). *Ç’anuris gramat’ik’uli analizi t’ekst’ebiturt*. Tbilisi, Georgia: Mecniereba.

Demirok, Ö. F. (2013). *Agree as a unidirectional operation: Evidence from Laz* (Unpublished master’s thesis). Boğaziçi University, İstanbul, Turkey.

Dixon, R. (1987). *Studies in ergativity*. Amsterdam, Netherlands: North-Holland.

Eren, Ö. (2016). *Spatial prefixes of Pazar Laz: A nano-syntactic approach* (Unpublished master’s thesis). Boğaziçi University, İstanbul, Turkey.

Fallon, P. D. (2016). *The synchronic and diachronic phonology of ejectives*. New York, NY: Routledge. doi:10.4324/9781315023809

Feurstein, W. (1983). *Untersuchungen zur materiellen kultur der Lazen* (Unpublished master’s thesis). Freiburg University, Freiburg, Germany.

Gerstenberger, C., Partanen, N. & Rießler, M. (2017). Instant annotations in ELAN corpora of spoken and written Komi, an endangered language of the Barents Sea region. In Arppe, A., Good, J., Hulden, M., Lachler, J., Palmer, A., & Schwartz, L. (Eds.), *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages* (pp. 57–66). Association for Computational Linguistics (ACL).

Grawunder, S., Simpson, A. & Khalilov, M. (2010). Phonetic characteristics of ejectives - samples from Caucasian languages. In S. Fuchs, M. Toda & M. Zygis (Eds.), *Turbulent Sounds* (pp. 209-244). Berlin: De Gruyter Mouton. doi:10.1515/9783110226584.209

Gürpınar, T. (2000). *The dialect of Pazar Laz and its case system* (Unpublished master’s thesis). Boğaziçi University, İstanbul, Turkey.

Haig, G. & Khan, G. (Eds.). (2018). *The languages and linguistics of western Asia*. Berlin, Boston: De Gruyter Mouton. doi:10.1515/9783110421682

Hammam, R. (2008). Indigenous languages of Indonesia: Creating language resources for language preservation. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages* (pp. 113–116). Asian Federation of Natural Language Processing.

Haznedar, B. (2018). The living Laz project: The current status of the Laz language and Laz-speaking communities in Turkey.

Hewitt, B. (2006). Georgia: Language situation. In K. Brown (Ed.), *Encyclopedia of language linguistics* (pp. 38–40). Elsevier. doi:10.1016/B0-08-044854-2/01797-1

Holisky, D. A. (1991). Laz. In A. C. Harris (Ed.), *The Kartvelian languages* (pp. 395–472). New York, NY: Delmar.

Hopcroft, J. E. & Ullman, J. D. (1969). *Formal languages and their relation to automata*. Boston, MA: Addison-Wesley Longman Publishing Co., Inc.

Hulden, M. (2009). Foma: A finite-state compiler and library. In Kreutel, J. (Ed.), *Proceedings of the Demonstrations Session at EACL 2009* (pp. 29–32). Association for Computational Linguistics (ACL).

Karttunen, L. & Beesley, K. R. (2001). A short history of two-level morphology. Presented at *ESSLLI-2001 Special Event titled "Twenty Years of Finite-State Morphology"*. Helsinki, Finland.

Karttunen, L., Chanod, J. P., Grefenstette, G. & Schiller, A. (1997). Regular expressions for language engineering. *Natural Language Engineering*, *2*(4), 305-328. doi:0.1017/S1351324997001563

Kavaklı, N. (2015). Novus ortus: The awakening of Laz language in Turkey. *İdil Journal of Art and Language*, *4*(16), 133–146. doi:10.7816/idil-04-16-08

Kavaklı, N. (2017). Language choice, use and transmission: Laz at the crossroads. *Journal of Endangered Language (Tehlikedeki Diller Dergisi)*, *7*(11), 51–66.

Kingston, J. (1985). *The phonetics and phonology of the timing of oral and glottal events* (Unpublished doctoral dissertation). University of California, Berkeley, CA.

Kirchhoff, K., Vergyri, D., Bilmes, J., Duh, K. & Stolcke, A. (2006). Morphology-based language modeling for conversational Arabic speech recognition. *Computer Speech & Language*, *20*(4), 589–608. doi:10.1016/j.csl.2005.10.001

Koehn, P. & Hoang, H. (2007). Factored translation models. In Eisner, J. (Ed.), *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 868–876). Association for Computational Linguistics (ACL).

Kondratyuk, D. (2019). Cross-lingual lemmatization and morphology tagging with two-stage multilingual BERT fine-tuning. In Nicolai, G., & Cotterell, R. (Eds.), *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology,* (pp. 12–18). Association for Computational Linguistics. doi:10.18653/v1/W19-4203

Koskenniemi, K. (1983). *Two-level morphology: A general computational model for word-form recognition and production*. Helsinki, Finland: University of Helsinki, Department of General Linguistics.

Kutscher, S. (2008). The language of the Laz in Turkey: Contact-induced change or gradual language loss? *Turkic languages, 12*, 82-102

Kutscher, S. (2011). On the expression of spatial relations in Ardeşen-Laz. *Linguistic Discovery, 9*(2). doi:10.1349/ps1.1537-0852.a.394

Kutscher, S. (2001). *Nomen und nominales syntagma im Lasischen. Eine deskriptive analyze des dialekts von Ardeşen*. München, Germany: Lincom Europa.

Lacroix, R. (2009). *Description du dialecte laze d'Arhavi (caucasique du sud, Turquie): grammaire et textes* (Doctoral dissertation). Retrieved from Networked Digital Library of Theses & Dissertations. (edsndl.oai.union.ndltd.org.theses.fr.2009LYO20091)

Lacroix, R. (2018). 6.2. Laz. In G. Khan & G. Haig (Eds.), *The languages and linguistics of western Asia* (pp. 830–860). Berlin, Germany: De Gruyter Mouton. doi:10.1515/9783110421682-021

Lazoğlu, F. & Feurstein, W. (1984). Lazuri alfabe. Laz alphabet. Entwurf eines Lazischen alphabetes, 160–161.

Linden, K., Silfverberg, M., Axelson, E., Hardwick, S. & Pirinen, T. (2011). HFST–framework for compiling and applying morphologies. In C. Mahlow & M. Pietrowski (Eds.), *Systems and frameworks for computational morphology* (pp. 67–85). Berlin: Springer.

Marr, N. (1910). *Grammatika çanskago Jazyka (the grammar of Laz)*. S. Petersburg, Russia: Academija.

Meurer, P. (2009). A computational grammar for Georgian. In P. Bosch, D. Gabelaia & J. Lang (Eds.), *Logic, Language, and Computation: 7th International Tbilisi Symposium on Logic, Language, and Computation, TbiLLC 2007, Tbilisi, Georgia* (pp. 1–15). Berlin, Germany: Springer.

Moseley, C. (Ed.). (2010). Atlas of the world's languages in danger (3rd ed.) [Interactive online edition]. Retrieved from http://www.unesco.org/languages-atlas/en/atlasmap/language-id-1056.html

Nivre, J., Agić, Ž., Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., & Augustinus, L. (2017). Universal dependencies 2.1.

Onal, E. & Tyers, F. (2019). Building a morphological analyzer for Laz. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 869–877. doi:10.26615/978-954-452-056-4_101

Öztürk, B. (2019a). *Applicatives in Pazar Laz* [South Caucasian Chalk Circle].

Öztürk, B. & Pöchtrager, M. A. (Eds.). (2011). *Pazar Laz*. München, Germany: LINCOM: Languages of the World Materials.

Öztürk, B. (2013). Low, high and higher applicatives: Evidence from Pazar Laz. In V. Camacho-Taboada, A. L. Jiménez-Fernández, J. Martín-González, & M. Reyes-Tejedor (Eds.), *Information Structure and Agreement* (pp. 275–295). Amsterdam, Netherlands: John Benjamins Publishing Company.

Öztürk, B. (2019b). The loss of case system in Ardeshen Laz and its morphosyntactic consequences. *STUF - Language Typology and Universals*, *72*(2), 193–219. doi:10.1515/stuf-2019-0008

Öztürk, B. (2021). Transitive unergatives in Pazar Laz. *Glossa: A Journal of General Linguistics*, *6*(1). doi:10.5334/gjgl.828

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137. doi:10.1108/eb046814

Pylkkänen, L. (2008). *Introducing arguments*. Cambridge, MA: MIT Press.

Sak, H., Saraclar, M. & Güngör, T. (2010). Morphology-based and sub-word language modeling for Turkish speech recognition. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing,* 5402–5405.

Salminen, T. (2007). Europe and North Asia. In C. Moseley (Ed.), *Encyclopedia of the world's endangered languages* (pp. 211–281). London, United Kingdom: Routledge. doi:10.4324/9780203645659.ch3

Sarı, İ. (2017). *Dil tarih ve gelenekleriyle Lazlar: Anadolu'da, Laz terimi, Karadeniz bölgesinde yaşayan bütün grupları ifade eden ortak bir addır.* İstanbul, Turkey: Noktaekitap.

Sennrich, R., Haddow, B. & Birch, A. (2016). Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1715–1725. doi:10.18653/v1/P16-1162

Singla, K., Sachdeva, K., Bangalore, S., Sharma, D. M. & Yadav, D. (2014). Reducing the impact of data sparsity in statistical machine translation. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation,* 51–56. doi:10.3115/v1/W14-4006

Sproat, R. & Stethem, S. (1992). *Morphology and computation*. Cambridge, MA: MIT Press.

Taylan, E. & Öztürk, B. (2014). Transitivity in Pazar Laz. *Acta Linguistica Hungarica*, *61*(3), 271–296. doi:10.1556/aling.61.2014.3.2

Türk, U., Bayar, K., Özercan, A. D., Öztürk, G. Y. & Özateş, S¸ . B. (2020). First steps towards universal dependencies for Laz. *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, 189–194.

Bucak'lişi, İ. A., Uzunhasanoğlu, H., & Aleksiva, İ. (2007). *Büyük Lazca sözlük: Didi Lazuri nenapuna*. İstanbul, Turkey: Chiviyazıları Yayınevi.