

University of Kentucky

UKnowledge

---

Theses and Dissertations--Linguistics

Linguistics

---


2024

## A Computer-Assisted Approach to Lexical Borrowing in Northeast Caucasian Languages

Bonnie Eleanor Wren-Hardin

*University of Kentucky*, [elliewh@gmail.com](mailto:elliewh@gmail.com)

Author ORCID Identifier:

 <https://orcid.org/0009-0001-7977-4011>

Digital Object Identifier: <https://doi.org/10.13023/etd.2024.125>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

### Recommended Citation

Wren-Hardin, Bonnie Eleanor, "A Computer-Assisted Approach to Lexical Borrowing in Northeast Caucasian Languages" (2024). *Theses and Dissertations--Linguistics*. 60.  
[https://uknowledge.uky.edu/lit\\_etds/60](https://uknowledge.uky.edu/lit_etds/60)

This Master's Thesis is brought to you for free and open access by the Linguistics at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Linguistics by an authorized administrator of UKnowledge. For more information, please contact [UKnowledge@lsv.uky.edu](mailto:UKnowledge@lsv.uky.edu).

## **STUDENT AGREEMENT:**

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

## **REVIEW, APPROVAL AND ACCEPTANCE**

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Bonnie Eleanor Wren-Hardin, Student

Dr. Andrew Byrd, Major Professor

Dr. Kevin B. McGowan, Director of Graduate Studies

A COMPUTER-ASSISTED APPROACH TO LEXICAL BORROWING IN  
NORTHEAST CAUCASIAN LANGUAGES

---

THESIS

---

A thesis submitted in partial fulfillment of the  
requirements for the degree of Master of Arts in the  
College of Arts and Sciences  
at the University of Kentucky

By

Bonnie Eleanor Wren-Hardin

Lexington, Kentucky

Director: Dr. Andrew Byrd, Professor of Linguistics

Lexington, Kentucky

2024

Copyright © Bonnie Eleanor Wren-Hardin 2024  
<https://orcid.org/0009-0001-7977-4011>

## ABSTRACT OF THESIS

### A COMPUTER-ASSISTED APPROACH TO LEXICAL BORROWING IN NORTHEAST CAUCASIAN LANGUAGES

*The disambiguation of loanwords and cognates can be a challenge, especially in areas where there has been intense language contact over an extended period of time, when the contact is between genetically related languages, and when the number of languages involved is large. Over the past several decades, more and more computational approaches to automatic cognate and borrowing detection have been created in an attempt to ease the load of examining hundreds to thousands of individual lexemes, as well as determine language family relationships with allegedly greater accuracy. While these methods are not perfect and cannot replace the knowledge or skillset of a linguist, this paper seeks to apply a computer-assisted, as opposed to purely computational, approach to lexical borrowing detection to three Northeast Caucasian languages spoken in a cluster of villages in Dagestan: Avar, Lak, and Archi. In this thesis, I utilize computational methods for cognate detection as a starting point, as well as a lexical distribution approach to borrowing, followed by qualitative methods for determining loanwords from borrowings as applied to the output of the computational methods.*

**KEYWORDS:** Northeast Caucasian, Language Contact, Lexical Borrowing, Dagestan.  
Computational Methods.

---

Bonnie Eleanor Wren-Hardin

---

4/25/2024

Date

A COMPUTER-ASSISTED APPROACH TO LEXICAL BORROWING IN  
NORTHEAST CAUCASIAN LANGUAGES

By  
Bonnie Eleanor Wren-Hardin

Dr. Andrew Byrd

---

Director of Thesis

Dr. Kevin B. McGowen

---

Director of Graduate Studies

4/25/2024

---

Date

## DEDICATION

To my parents ☺

## ACKNOWLEDGMENTS

I am so grateful to the many people who helped this thesis come to life!

First, I would like to thank my committee chair, Dr. Andrew Byrd, who has been the best chair I could have asked for. Thank you for all of the questions, advice, and biweekly organizational meetings!

I would also like to thank my committee members: Dr. Rusty Barrett, Dr. Josef Fruehwald, and Dr. Mark Lauersdorf. The three of you as well as Dr. Byrd have spent a remarkable amount of time meeting with me and providing direction as well as resources, which I so appreciate. All of your perspectives have greatly informed the direction of my work for the better.

My time at the University of Kentucky has been so informative, and I would like to thank all of the other professors I have had the privilege of taking courses with. The work I have undertaken as a part of this program has expanded my knowledge and shaped who I am (and who I hope to be) as a linguist.

I am deeply grateful to all the graduate students of the MALTT program for their friendship and support. A special thank you must be extended to Nour and Chase (the Oklahoma crew!), as well as Catie, John, and Connor, for being great friends throughout it all. I am also thankful in particular for the support of the rest of my cohort not yet named: Iain, Angel, and Ian.

My professors at the University of Oklahoma were influential in setting me on my path; a special thank you must be extended as well in this case to Dr. Mark Norris, for assigning the work that sparked my interest in the Caucasus as well as encouraging me to

always ask questions and investigate deeply and thoroughly. The fact that this thesis is on this topic is in large part because of your Typology course all those years ago!

Thank you also to my partner, Kevin, for your persistent love and support, as well as my cat, Kilgore, who spent a lot of time sitting cutely in my lap while I was writing this.

Lastly, to my family and in particular my parents, I am forever grateful for all of the support you have given me, not just during this program, but throughout my whole life. Thank you for always believing in me and pushing me to achieve my own goals.



## TABLE OF CONTENTS

ACKNOWLEDGMENTS .....	iii
LIST OF TABLES .....	vii
LIST OF FIGURES .....	ix
CHAPTER 1. INTRODUCTION .....	1
1.1 Introduction .....	1
CHAPTER 2. THEORETICAL, METHODOLOGICAL, AND CONTEXTUAL BACKGROUND .....	3
2.1 Comparative Method .....	3
2.1.1 Contact-Driven Change .....	6
2.1.2 Northeast Caucasian Language Family .....	8
2.1.2.1 Patterns of language contact and multilingualism in the Northeast Caucasian language family .....	11
2.1.2.2 External Sources of Language Contact .....	16
2.1.2.3 Villages of Focus .....	18
2.1.3 Northeast Caucasian and the Comparative Method .....	23
2.1.4 Northeast Caucasian and Lexical Borrowing Studies .....	25
2.2 Computational Methods for Identification of Cognates and Loanwords .....	27
2.2.1 Data Sources for Computational Methods .....	28
2.2.2 Automatic Cognate Detection .....	29
2.2.2.1 LingPy .....	31
2.2.3 Automatic Borrowing Detection .....	37
2.2.4 Automated versus Computer-Assisted Approaches .....	39
2.2.5 Computation Methods Applied to the Caucasus .....	42
CHAPTER 3. DATA AND METHODS OF THE PRESENT STUDY .....	43
3.1 Data .....	43
3.1.1 DagSwadesh .....	44
3.1.2 Intercontinental Dictionary Series .....	45
3.1.3 Atlas of Multilingualism in Dagestan .....	46
3.2 Computational Methods .....	47
3.2.1 Evaluating Cognate Detection Methods .....	48
3.2.2 Applying the Selected Method to the Village Cluster .....	50
CHAPTER 4. RESULTS .....	53
4.1 Testing on DagSwadesh .....	53
4.2 Implementation on the Intercontinental Dictionary Series .....	56
4.2.1 Transcription Style .....	70
4.2.2 Possible Borrowings from Avar .....	70
4.2.3 Possible Borrowings from Lak .....	95
CHAPTER 5. SUMMARY .....	102

5.1	<i>Borrowings in Context</i> .....	102
5.2	<i>Conclusion</i> .....	105
	REFERENCES .....	107
	VITA.....	114

## LIST OF TABLES

Table 1: The SCA Sound Class Model (List, 2012b, p. 43).....	33
Table 2: “Relevant patterns of distribution of lexically similar forms in languages of the region, and the corresponding borrowing or inheritance history of such a form” (Moro et al., 2023, p. 220).....	41
Table 3: Possible distribution patterns of lexical items and hypothesized explanations...51	51
Table 4: SCA Method Precision, Recall, and F-Scores.....	53
Table 5: LexStat Method Precision, Recall, and F-Scores.....	54
Table 6: Possible distribution patterns of lexical items and hypothesized explanations...58	58
Table 7: Overall Results.....	59
Table 8: Set associated with 'navel'.....	61
Table 9: Set associated with 'spring'.....	62
Table 10: Example of an undetected internal loan.....	62
Table 11: Results of Hypothesis 1.....	65
Table 12: Results of Hypothesis 2.....	66
Table 13: Results of Hypothesis 3.....	67
Table 14: Results of Hypothesis 4.....	68
Table 15: Results of Hypothesis 5.....	69
Table 16: Transcription Conventions.....	70
Table 17: Set associated with 'booty, spoils'.....	71
Table 18: Set associated with 'heart'.....	72
Table 19: Set associated with 'stinking, bad-smelling'.....	72
Table 20: Set associated with 'Friday'.....	73
Table 21: Set associated with 'arrow'.....	73
Table 22: Set associated with 'bow'.....	74
Table 23: Set associated with 'bark'.....	75
Table 24: Set associated with 'blister'.....	75
Table 25: Set associated with 'blood'.....	76
Table 26: Set associated with 'boundary'.....	78
Table 27: Set associated with 'cattle'.....	78
Table 28: Set associated with 'cock, rooster'.....	79
Table 29: Set associated with 'daughter-in-law'.....	79
Table 30: Set associated with 'ditch'.....	80
Table 31: Set associated with 'east'.....	80
Table 32: Set associated with 'family'.....	81
Table 33: Set associated with 'father-in-law'.....	81
Table 34: Set associated with 'fermented drink'.....	82
Table 35: Set associated with 'fisherman'.....	82
Table 36: Set associated with 'forehead'.....	83
Table 37: Set associated with 'hammer'.....	83

Table 38: Set associated with ‘idol’ .....	84
Table 39: Set associated with ‘island’ .....	84
Table 40: Set associated with ‘keep, retain’ .....	85
Table 41: Set associated with ‘leather’ .....	85
Table 42: Set associated with ‘lion’ .....	86
Table 43: Set associated with ‘lip’ .....	86
Table 44: Set associated with ‘live, living, life’ .....	87
Table 45: Set associated with ‘magic, witchcraft, sorcery’ .....	87
Table 46: Set associated with ‘mosquito’ .....	88
Table 47: Set associated with ‘mother-in-law’ .....	89
Table 48: Set associated with ‘oar’ .....	89
Table 49: Set associated with ‘plaintiff’ .....	90
Table 50: Set associated with ‘queen’ .....	90
Table 51: Set associated with ‘ring (for finger)’ .....	91
Table 52: Set associated with ‘sea’ .....	91
Table 53: Set associated with ‘sorcerer, witch’ .....	92
Table 54: Set associated with ‘suspect’ .....	92
Table 55: Set associated with ‘taste’ .....	93
Table 56: Set associated with ‘thief’ .....	93
Table 57: Set associated with ‘tribe, clan’ .....	94
Table 58: Set associated with ‘widow’ .....	94
Table 59: Set associated with ‘yesterday’ .....	95
Table 60: Set associated with ‘army’ .....	95
Table 61: Set associated with ‘bull’ .....	96
Table 62: Set associated with ‘eyebrow’ .....	96
Table 63: Set associated with ‘frog’ .....	97
Table 64: Set associated with ‘glove’ .....	97
Table 65: Set associated with ‘lake’ .....	98
Table 66: Set associated with ‘owl’ .....	99
Table 67: Set associated with ‘root’ .....	99
Table 68: Set associated with ‘short’ .....	100
Table 69: Set associated with ‘snow’ .....	101

## LIST OF FIGURES

Figure 1: Caucasus: Administrative division (Koryakov, 2020).....	9
Figure 2: Nakh-Daghestanian Languages (Ganenkov & Maisak, 2020, p. 88).....	11
Figure 3: Map of the Archib Cluster (Dobrushina, 2013) .....	19
Figure 4: Alignment analysis of German Tochter and English daughter (List, 2012b) ....	29
Figure 5: Modeling the directionality of sound change patterns in scoring schemes (List, 2012b, p. 40).....	33
Figure 6: SCA Workflow [based on List (2012b, p. 42)].....	34
Figure 7: SCA distance versus LexStat distance (List, 2012a, p. 122) .....	36
Figure 8: Example of computing the B-Cubed precision and recall for one item (Amigó et al., 2009, p. 471).....	49
Figure 9: SCA Results .....	54
Figure 10: LexStat Results .....	55
Figure 11: Multilingualism over time in Archib (Dobrushina et al., 2017) .....	104

## CHAPTER 1. INTRODUCTION

### 1.1 Introduction

The disambiguation of loanwords and cognates can be a challenge, especially in areas where there has been intense language contact over an extended period of time, when the contact is between genetically related languages, and when the number of languages involved is large. Separating loanwords from cognates allows for the possibility of reconstructing several important pieces of information: the genetic history of the language, the proto-language (or ancestor language), and contact situations of the past. The genetic history of the language demonstrates how the language family split and diverged over time, as well as which languages are most closely related to one another, providing critical information needed to begin the reconstruction process of the proto-language. On the other hand, identifying loanwords can inform us of social and migratory connections that may have existed and are now gone. However, as both genetic inheritance and contact leads to increased similarity in the linguistic systems of the languages involved (Epps et al., 2013, p. 213), performing such analysis can be challenging.

Over the past several decades, computational approaches to historical linguistics, and in particular automatic cognate and borrowing detection, have begun to proliferate. These approaches have the potential to ease the load of examining hundreds to thousands of individual lexemes, as well as determine language family relationships with allegedly greater accuracy. However, these methods are not perfect, and are unable to replicate the skills and knowledge of trained historical linguistics; as such, they may function best as “triage” programs, bringing to light interesting cases and information that can benefit from

a more rigorous analysis. As a result, the idea of a “computer-assisted”, as opposed to purely computational, approach to historical linguistics has also begun to gain traction. This paper seeks to apply a computer-assisted approach to lexical borrowing detection to three Northeast Caucasian languages spoken in a cluster of villages in Dagestan: Avar, Lak, and Archi.

The Northeast Caucasian language family is the perfect test-case for such a computer-assisted approach as the family is large, with 40+ lects, as well as densely distributed in Dagestan. Historical patterns of small-scale multilingualism means that there has been consistent contact between related languages for centuries. The end result of these circumstances is that there has been intense language contact in the region and among speakers of these languages for thousands of years, leading to difficulty in separating true cognates from borrowings.

In this thesis, I utilize computational methods for cognate detection as a starting point, as well as a lexical distribution approach to borrowing. Then, I apply qualitative methods for determining loanwords from borrowings to the output of the computational methods.

## CHAPTER 2. THEORETICAL, METHODOLOGICAL, AND CONTEXTUAL BACKGROUND

### 2.1 Comparative Method

The process of identifying genetic relationships between languages is essential to understanding the historical development of said languages and the language families they are a part of. Frequently no written records exist for an ancestor language, which becomes more true the further back in time the set of languages in question diverged (Campbell, 2013, p. 109). Therefore, the existence of the daughter languages themselves, as well as possibly written records of earlier stages in the development of the daughter languages, may be the only information available for the reconstruction of proto-languages and the determining of genetic relationships between the daughter languages.

Historically, the Comparative Method has been the method most utilized to identify and determine these genetic relationships and reconstruct proto-languages, which are the hypothesized ancestor languages from which a modern language or group of modern languages are assumed to have descended (Campbell, 2013, p. 107). The Comparative Method is a process by which words in a set of daughter languages are systematically compared with one another in order to determine regular patterns of sound change (Campbell, 2013, p. 107; Ross & Durie, 1996, pp. 6-7). These sound change patterns can then be used to identify the sounds of the earlier stages of the language family and systematically reconstruct words in the proto-language. Underlying the Comparative Method is the idea that all languages undergo regular, internal sound change (Campbell, 2013, p. 111).

Critically, the Comparative Method is an iterative and, frequently, time-consuming process. While various linguists may divide or describe some steps differently, the goals



and actions of the Comparative Method are the same or similar across sources. Ross and Durie (1996, pp. 6–7) describe their steps as follows:

1. Determine on the strength of diagnostic evidence that a set of languages are genetically related, that is, that they constitute a ‘family’;
2. Collect putative cognate sets for the family (both morphological paradigms and lexical items).
3. Work out the sound correspondences from the cognate sets, putting ‘irregular’ cognate sets on one side;
4. Reconstruct the protolanguage of the family as follows:
  - a. Reconstruct the protophonology from the sound correspondences worked out in (3), using conventional wisdom regarding the directions of sound changes.
  - b. Reconstruct protomorphemes (both morphological paradigms and lexical items) from the cognate sets collected in (2), using the protophonology reconstructed in (4a).
5. Establish innovations (phonological, lexical, semantic, morphological, morphosyntactic) shared by groups of languages within the family relative to the reconstructed protolanguage.
6. Tabulate the innovations established in (5) to arrive at an internal classification of the family, a ‘family tree’.
7. Construct an etymological dictionary, tracing borrowings, semantic change, and so forth, for the lexicon of the family (or of one language of the family).

While the steps appear to proceed linearly above, in reality, and as described by Ross and Durie (1996, p. 7), the steps involved are in practice completed in a recursive manner, with new data and information constantly necessitating the repetition of various steps. To even start the process, one must have languages they suspect to be related or to form a family, implying they may have already identified suspected or purported cognate sets. As another example of the iterative nature of the process, after completing steps one-four, more purported cognate sets might be found, leading to revisions in the sound correspondences. Alternatively, other languages might be proposed as members of the family, leading to the development of new cognate sets and a new family tree in step six. Even a single step, such as step three, “Work out the sound correspondences from the cognate sets, putting ‘irregular’ cognate sets on one side” (Ross & Durie, 1996, p. 7), requires repetition. As new sound correspondences are identified, previously identified ones may need to be revised or changed. Essentially, the Comparative Method is an iterative process, requiring many steps to be repeated many times, with the results being constantly modified and updated.

Additionally, the amount of time needed to gather and compare cognate sets across large language families also protracts the process. The Austronesian language family, for example, is made up of approximately 1,200 languages (Kikusawa, 2015, p. 657); attempting to compare even a single set of reflexes across 1,200 languages would be incredibly time-consuming and onerous to complete manually. Establishing regular sound correspondences also requires far more than a single reflex, meaning one might need to

compare tens of thousands of reflexes to be able to feel as though they are able to solidly propose reconstructions of morphemes and a reliable family tree.

Lastly, the results of the Comparative Method function purely as a hypothesis that cannot be proven or disproven (Campbell, 2013, pp. 127–128). While the Comparative Method is considered methodologically strong and has been applied to language families all over the world, it developed primarily out of application to the Indo-European language family. The Indo-European language family is “lucky” in that many (although obviously not all) branches contain written records documenting earlier stages in the development of the languages (Campbell, 2013, p. 109). The existence of these written records allows for even earlier stages of the language families to be verified and compared with one another, and they provide key information regarding sound changes that have already occurred. However, the vast majority of the approximately 7,000 currently spoken languages of the world do not have historical written records dating back as far as those from the Indo-European language family, or even any written records at all from any time period (Campbell, 2013, p. 109). In these cases, the currently existing daughter languages are the only records of these proto-languages, complicating the process of reconstructing their proto-languages.

### 2.1.1 Contact-Driven Change

In addition to the logistical difficulties presented by the Comparative Method itself, the method also emphasizes genetic relationships and vertical transmission, ignoring or perhaps underrepresenting the effects of language contact and horizontal transmission. The emphasis on regular, internal sound changes identified through sound correspondences strongly promotes a model of language in which languages are discrete items each

stemming perfectly from an ancestor language with little outside interference. In reality languages do not exist in isolation, and all languages are influenced through contact with other languages to some degree.

Contact driven change can have effects at every level of the linguistic system, including within the lexicon, the phonological system, and the morpho-syntax (Thomason, 2001, p. 69; Winford, 2003). Language contact is perhaps most immediately visible when occurring between languages that are not genetically related; however, languages that are closely related can and do still influence each other as a result of contact, leading to difficulties in distinguishing between contact effects and genetic inheritance (Epps et al., 2013, p. 213). Additionally, not all kinds of contact affect the linguistic system equally, and the actual linguistic outcomes of language contact are influenced by a variety of factors, including the socio-cultural, economic and political contexts of the speech communities in contact, as well as the typological similarities existing between the languages (Sankoff, 2004; Thomason & Kaufman, 1988). For example, the degree and directionality of multilingualism, as well as the age of second language learning, can affect degree and directionality of borrowing and contact effects, as well as purported complexification or simplification within the morphosyntactic system (Trudgill, 2010). From the point of view of historical linguistics, such contact situations can lead to issues with determining subgroupings of language families (Epps et al., 2013, p. 211; Pat-El, 2013, p. 314).

The lexicon is often perceived as one of the easiest and earliest places for contact effects to appear, with lexical borrowing possible even at the lowest level of contact in Thomason's borrowing scale (2001, pp. 70-71). The identification of loanwords is then an

important step in the Comparative Method for the purpose of identifying regular sound correspondences, reconstructing proto-languages, and identifying the genetic relatedness of languages, but can also be informative for reconstructing contact situations between language groups and developing a typology of language contact outcomes. Loanwords must be removed from potential cognate sets in order to avoid creating false reconstructions or determining false genetic relations between languages. Loanwords can be identified in a variety of ways, including their phonology, morphology, absence or presence in cognate sets, semantics, and the geographic, ecological, and cultural factors of the languages in question (Campbell, 2013, pp. 62–66).

Lexical borrowings may be easier to identify when the borrowed words are from a different language family than the recipient language, as the original items may have phonemes or phonological properties that are not present in the recipient language that must be altered in predictable ways, or the sister languages of the recipient language may not have a similar form in their language, indicating borrowing from another language. However, when borrowing is between related languages, it can be more complicated to separate inherited vocabulary from borrowed vocabulary, providing another layer of difficulty in establishing cognate sets for the Comparative Method.

### 2.1.2 Northeast Caucasian Language Family

A language family that exemplifies all of the above described difficulties is the Northeast Caucasian language family, spoken in the Caucasus. The Caucasus is a region containing a crest mountain range subdividing Asia and Europe that can be divided into the North Caucasus, roughly containing the republics of Adyghe, Dagestan (also written as Daghestan), Chechnya, Ingushetia, North Ossetia-Alania, Kabardino-Balkaria, and



(Amiridze, 2019; Comrie, 2008; Dobrushina et al., 2020b; Tuite, 1999). For the remainder of the discussion, I will use the terms South Caucasian, Northwest Caucasian, and Northeast Caucasian, respectively.

Even within this linguistically diverse region, the republic of Dagestan, with a population of three million (Facts about Russia's Republic of Dagestan | Reuters, 2023) is notably diverse (Comrie, 2008, p. 134), with over forty Northeast Caucasian languages being spoken in the republic alone (Dobrushina et al., 2020b, p. 54). The Northeast Caucasian language family was historically thought to be split into two branches, Nakh and Daghestanian, which was the origin of the earlier language family name “Nakh-Daghestanian” (Dobrushina et al., 2020b, p. 30). Now, however, the language family is usually split more evenly into several sub-branches (Dobrushina et al., 2020b; Ganenkov & Maisak, 2020; Schulze, 2017). Schulze (2017, pp. 107–108) proposes five branches: Nakh, Avar-Andic, Tsezic, Lako-Dargi, and Lezgian. Dobrushina et al (2020b, pp. 30–32) propose six branches: Nakh, Avar, Andic, Tsezic, Dargi, and Lezxic. Ganenkov and Maisak (2020, p. 88) alternatively propose four branches, Nakh, Avar-Andic-Tsezic, Dargwa, and Lezxic, and two “family-level isolates”, Lak and Khinalug. While there are questions as to the placement of individual languages (see Dobrushina (2020b) for discussion of Lak and Khinalug, as well as Kassian and Testelelets (2017) for Hinuq), most differences seem to be concerned with whether to split or combine the Avar, Andic, and Tsezic branches into individual branches or a single, larger branch. Ganenkov and Maisak (2020, p. 88) combine them into the Avar-Andic-Tsezic branch, Schulze (2017, pp. 107–108) combines Avar and Andic to form the Avar-Andic branch separate from the Tsezic branch, and Dobrushina et al (2020b, pp. 30–32) report separate Avar, Andic, and Tsezic

branches. Figure 2 demonstrating the branches of the language family as proposed by Ganenkov and Maisak (2020, p. 88) is below, and this is the version of the family tree that will be utilized throughout this thesis.

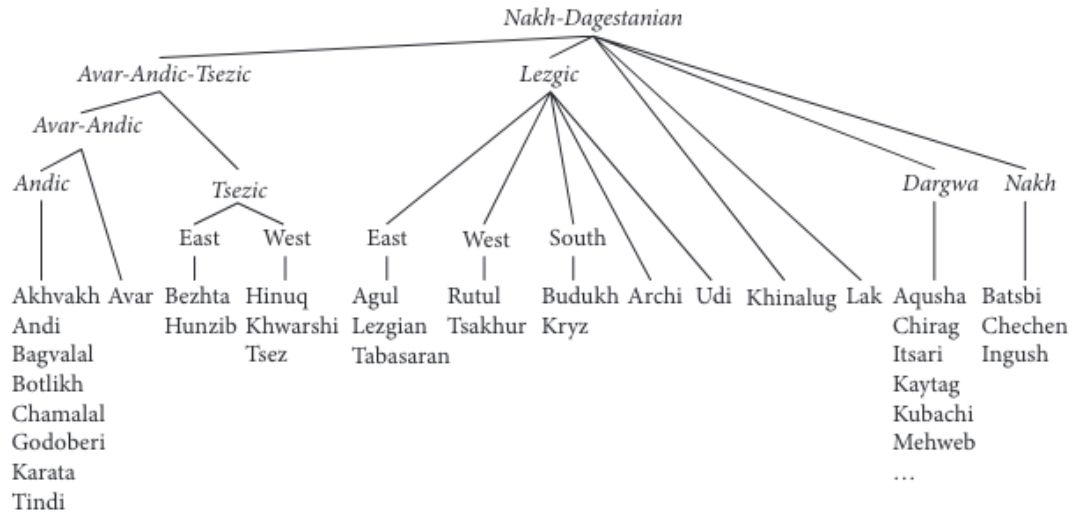


Figure 2: Nakh-Daghestanian Languages (Ganenkov & Maisak, 2020, p. 88)

Some of the Northeast Caucasian languages of Dagestan, such as Hinuq, have speaker counts in the hundreds, whereas most have speakers counts in the thousands (e.g., Bezhta, Chirag, and Archi) or tens of thousands (e.g., Rutul and Andi) (Dobrushina et al., 2020b, p. 31). Several, such as Avar and Lezgian, even have speaker counts in the hundreds of thousands (Dobrushina et al., 2020b, p. 31). Before the promotion of Russian in the last one hundred years, Dagestan had been historically representative of a form of small-scale multilingualism in which villagers speaking the language of their community also speak the languages of the villages near them or languages that they need for trade, without an overarching lingua franca encompassing the entire region (Dobrushina, 2023).

#### 2.1.2.1 Patterns of language contact and multilingualism in the Northeast Caucasian language family



The number of languages spoken in such an area, the lack of a large-scale lingua franca prior to the promotion of the Russian language, as well as the economic and sociocultural ties in the region, has meant that there has been intense language contact between the languages of the Northeast Caucasian language family for thousands of years. The factors affecting the linguistic outcomes of language contact can vary, including geography, economy, sociocultural factors, language typology, levels and patterns of multilingualism, intensity of contact, and language ideologies. Additionally, many of these factors are interrelated, affecting one another; for example, the economy of a region can impact the intensity of contact and the resulting patterns of multilingualism. Therefore, the impacts of language contact on the linguistic system and the linguistic landscape are complex and intertwined with one another. Within the Caucasus region, and particularly in Dagestan, Nichols (2013) proposes a connection between geography, economy, and multilingualism that leads to asymmetrical borrowing and morphosyntactic complexity.

In mountainous regions, such the Caucasus, which is specifically a crest mountain, levels of multilingualism and of structural complexity can appear to pattern with the altitude at which the language is spoken, with instances of asymmetrical vertical bilingualism and a tendency towards morphosyntactic complexity at higher altitudes (Nichols, 2013; Urban, 2020). Such crest mountain geography of the Caucasus impacts the types and frequency of language contact that can be expected, with different patterns appearing for those in the highlands versus those in the lowlands based largely on the economic factors involved (Nichols, 2013, p. 39). The physical geography of the highlands in the mountains means that large, open networks and strong, independent economies are difficult to maintain; the lowlands are able to be more economically independent and

dominant due to their preferred geographic location allowing for longer and better crop growing seasons as well as access to Silk Road trade routes (Nichols, 2013, pp. 39, 43). Additionally, within the Caucasus and Dagestan in particular, highlanders have traditionally been transhumant, meaning they must move their livestock to pastures at lower altitudes in the winter months. Since the markets and pastures that these transhumant highlanders rely on are commonly in the lowlands (Nichols, 2013, p. 57; Urban, 2020, p. 4), contact between languages is typically, therefore, the result of economic necessity on the part of the highland language group (Dobrushina, 2013).

Language contact between groups with disparate amounts of economic power affects each language involved differently. The speaker group that is more economically dependent would likely have a greater need to speak the language of the group that they are reliant on, but not the other way around (Nichols, 2013, p. 43). This produces patterns of what is referred to as asymmetrical multilingualism, in which one group is able to speak the language of another group, but the second group is not able to speak the language of the first (Nichols, 2013, p. 43). Adding in the variable of altitude, the economic conditions created by the mountain geography of the Caucasus means that highlanders, who are more economically dependent, would by necessity speak the languages of the lowlanders, who had more stable economies as well as the pastures the highlanders relied on for their livestock in winter, but the lowlanders would not speak the languages of the highlanders (Nichols, 2013, p. 43). As a result, the lowland languages in the Caucasus acted as “local *lingue franche* on a vertical basis” (Urban, 2020, p. 5). Overall, geographic and economic factors have combined to create an asymmetrical pattern of multilingualism in the region

such that those in the highlands are likely to speak the language of the lowlanders below them, but not vice versa.

Additionally, asymmetrical multilingualism has impacts on the linguistic system that extend beyond the mere fact of multilingualism, such as language change, spread, and borrowing. The one-way, asymmetric nature of the multilingualism means that highlanders who speak lowland languages could likely incorporate lowland words and structures, whereas the reverse would be less likely to happen; as a result, there is an “upward spread of isoglosses, dialects, and languages” (Nichols, 2013, p. 58). This dialect and language shift may also be the result of the highlanders’ “secondary claim to essential economic resources” (Nichols, 2013, p. 57). Essentially, the geography of the mountain region promotes economic dependency and a lack of essential economic resources in the highlands, which leads to a pattern of asymmetrical vertical multilingualism; this asymmetrical vertical multilingualism, as well as the economic status of the highlands, leads to upwards isogloss, dialect, and language shift.

Geography can also lead to increased linguistic complexity or simplification by promoting economic and sociocultural factors that may affect the linguistic system. However, it is important to recall that the notion of linguistic complexity is itself a complex topic; definitions vary, typically focusing on morphosyntax, and often whether features are seen as more or less complex hinges on the specific lens through which the feature is being viewed. In the case of the Caucasus, some have argued that mountain geography can lead to linguistic complexification by creating the economic and sociolinguistic features necessary to conserve complexities in the highlands (Nichols, 2013; Urban, 2020). The asymmetrical vertical multilingualism of Dagestan, for example, means that the lowland

languages would have a higher number of adult second language speakers, which could lead to more morphosyntactic simplifications in the language (Trudgill, 2015, p. 145), whereas “smaller” languages with fewer native speakers often have far fewer adult second language speakers (Dobrushina & Moroz, 2021). These “smaller” languages are typically focused in the highlands, and consequently, the highland languages may preserve more complexities than the lowland languages due to their lower number of adult second language learners (Nichols, 2013, p. 39). Additionally, any new complexities generated would be more likely to be conserved in a highland language than a lowland language as well, since there are fewer adult second language learners. In general, the higher the altitude of the highland language, the more isolated it would be expected to be and the lower the number of adult second language learners it would be expected to have; therefore, the higher the altitude of the language, the more complexities it would be expected to preserve (Nichols, 2013).

More broadly, the outcomes of language contact are also influenced by the language ideologies and social practices of the region and individual speakers. While many communities with small-scale multilingualism are exogamous, meaning members intentionally marry outside of their language group, which can help to maintain multilingualism in smaller regions, it is also possible for endogamous language areas to be multilingual (Pakendorf et al., 2021, pp. 847-849). Dagestan is one such area, with clan and village-based endogamy being commonplace until at least the mid-21st century in villages and language groups of all sizes (Dobrushina, 2023). Even in rare cases in Dagestan involving exogamous marriages, it was always the women leaving their villages and being “married out” to the husband, and never the other way around (Dobrushina,

2023). Additionally, in these cases the women are expected to learn the language of their husband, and the children are raised with the father's language as their first language (Dobrushina et al., 2020b). Dobrushina (2023) argues that endogamous marriages were the norm in Dagestan due to concerns surrounding land inheritance; as villages in the highlands did not have a lot of land for pastures and farming, what did exist was critical and coveted.

Patterns of endogamy and exogamy are tied to implicit and explicit language ideologies; explicit language ideologies reflect what those say about the languages used by themselves and others, such as what language use is acceptable when and by who, while implicit language ideologies reflect unconscious beliefs, and can often be seen in the “contradictions between observed language use and explicit ideologies” (Pakendorf et al., 2021). In this case, the explicit ideologies revolve around expectations of endogamy within Dagestan language groups; the villages themselves are linguistically homogeneous as a result of the clan and village-based endogamy, but multilingualism is still pervasive due to interactions with neighboring villages for economic necessity (Dobrushina, 2023; Pakendorf et al., 2021).

#### 2.1.2.2 External Sources of Language Contact

While small-scale multilingualism is a prominent feature of the historical patterns of contact within Dagestan, there has been additional contact amongst Northeast Caucasian languages by more distant languages as well, often patterning with the differences between ‘internal’ and ‘external’ sources of contact. An ‘internal’ source of contact is a language from the same family as the language of reference, while an ‘external’ source of contact is a language from an alternate language family than the language of reference. Within the

context of Dagestan and the Northeast Caucasian language family, it has been possible for both internal and external languages to participate in either small-scale or distal multilingualism, depending on the language and village of focus.

Persian and Arabic are external and distal sources of contact in Dagestan historically. Persian contact was strongest in the southern part of Dagestan in the 3rd to 6th centuries (Dobrushina & Kultepina, 2021, p. 341) and was gone by at least the late Middle Ages (Daniel et al., 2021, p. 528). Most contact from Arabic is the result of the conversion of the majority of the population of the area to Islam from the 7th to 15th centuries (Dobrushina & Kultepina, 2021, p. 341). As a result, the use of Arabic was largely for religious purposes, and was “the language of prayer” (Dobrushina & Kultepina, 2021, p. 341).

Two other sources of external borrowing, Kumyk and Azerbaijani, both Turkic languages that are somewhat mutually intelligible (Wixman, 1980, p. 109), have participated in both small-scale and distal multilingualism in Dagestan, though distal multilingualism has been more common. Kumyk was spoken in the lowlands of Dagestan until approximately the 1850s (Dobrushina & Kultepina, 2021, p. 341), and speakers of the village of Rikvani in Dagestan indicated it was spoken by a small percentage of their village (18%) as a result of winter shepherding (Dobrushina et al., 2020a, p. 28). Azerbaijani was used within southern Dagestan, which borders Azerbaijan. In the past, when the border was nonexistent or more permeable, whole villages would leave Dagestan and spend the winter in villages in Azerbaijan, as there was greater access to resources and better weather (Chechuro et al., 2021, p. 1022).

The final external source of language contact of note in Dagestan is Russian. Russia began to have a larger presence in the region in the 16th century, and Dagestan was annexed to the Russian Empire in the 19th century (Dobrushina & Kultepina, 2021, p. 341). However, the Russian language did not play a large role in the region until the mid 20th century when “Russian teachers were sent to Daghestanian villages” and Russian began to be taught in schools (Dobrushina et al., 2019, p. 3). In the late 19th century, less than 1% of the population of Dagestan spoke Russian, while now that number is over 90% (Dobrushina & Kultepina, 2021, p. 339). As a result, Russian has begun to replace the historic patterns of small-scale multilingualism through functioning as a regional lingua franca (Dobrushina & Kultepina, 2021, p. 341).

#### 2.1.2.3 Villages of Focus

The villages in focus in this thesis, Archib, Shalib, and Chitab, are all located within walking distance of one another in the Charoda district of Dagestan, Russia (Dobrushina, 2013). Each village has a different primary native language, each from a different branch of the Nakh-Daghestanian language family. Archi (Lezgian branch), is spoken in Archib, dialectal Lak (Lak branch) is spoken in Shalib, and dialectal Avar (Avar-Andic-Tsezic branch) is spoken in Chitab (Dobrushina, 2013). These villages were selected due to their presence in the Atlas of Multilingualism of Dagestan (Dobrushina et al., 2017), meaning the patterns of multilingualism present in these villages over the past approximately 150 years has been thoroughly documented and described.

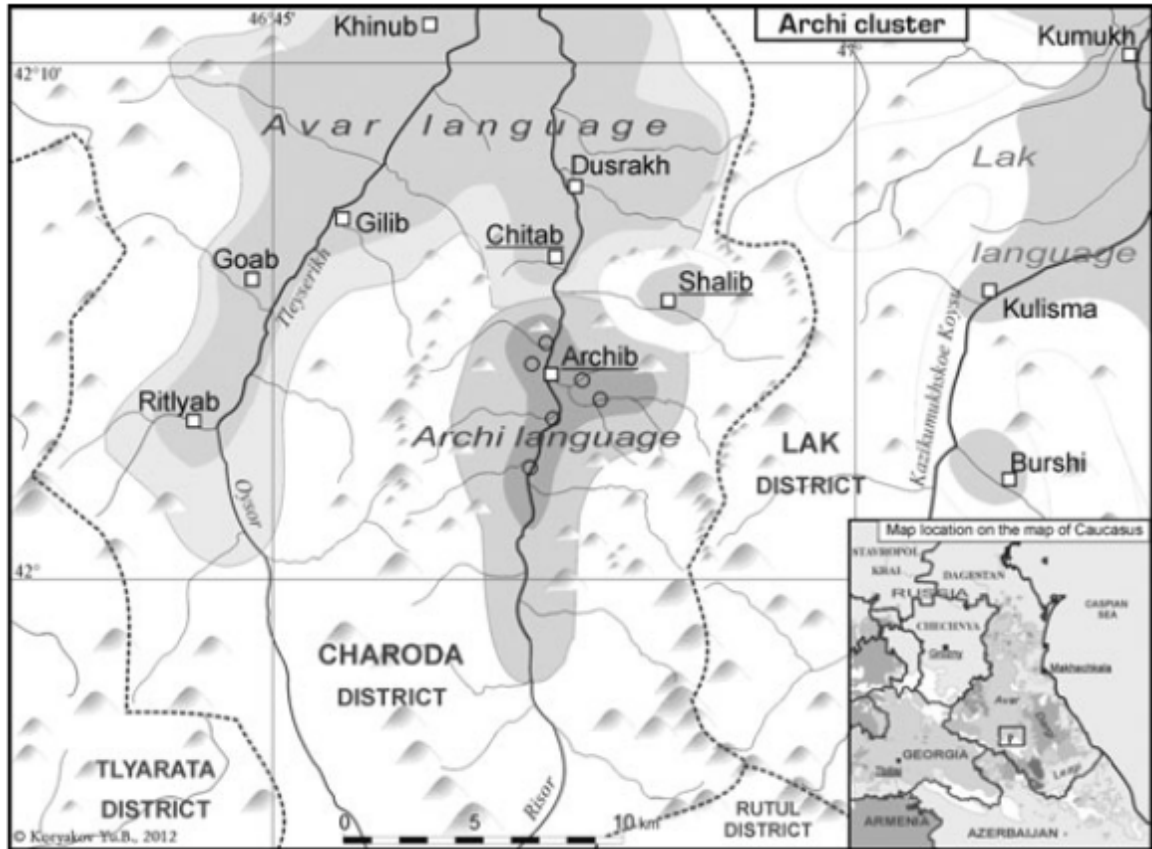


Figure 3: Map of the Archib Cluster (Dobrushina, 2013)

In addition to being from different branches of the Northeast Caucasian language family, each language and village also has markedly different sociolinguistic profiles, especially within the district in question. Avar is by far the most widely spoken indigenous language of the Caucasus in Dagestan, with a speaker population of approximately 850,000, or 30% of the population of Dagestan (Dobrushina, 2013), and is also one of 14 official languages of Dagestan (Forker, 2020). The first evidence of written Avar is from the 1400s, but it was not until the 1700s that an orthography for Avar was developed and texts were more widely created (Forker, 2020, p. 243). Avar itself can be divided into two main dialect groups, northern and southern (Forker, 2020). The village of Chitab speaks a dialect of Avar and is within proximity of other Avar-speaking villages, as the Avars



comprise 90% of the Charoda district, and children are taught both standard Avar and Russian in schools (Dobrushina, 2013).

The Lak speaker population within Dagestan is not as large as that of Avar, but is still quite sizable, at approximately 160,000 speakers, or 5.5% of the population (Friedman, 2020, p. 201). Lak is considered another one of the “major indigenous literary languages of Dagestan” (Dobrushina, 2013, p. 380) and has been a literary language since at least 924 CE (Friedman, 2020, p. 202). Lak can also be split into two additional branches of dialects (Friedman, 2020, p. 202). Unlike Chitab, which has neighboring villages sharing their first language of Avar, Shalib is the only village within the Charoda district that speaks Lak; the language is primarily spoken in the districts of Lak, Kuli, and Novolak (Dobrushina, 2013; Friedman, 2020). As a result, the Lak of Shalib are separated from other Lak speaking groups by the Shalib and Dulti mountain ranges, with the village of Archib as their closest neighbor (Friedman, 2020, p. 202). Shalib students learn both standard Lak and Russian in school (Dobrushina, 2013, p. 380).

Unlike Chitab and Shalib, which are single villages within much larger groups of Avar and Lak speakers respectively, Archi is spoken only by 1,200 people living in seven settlements within walking distance of each other, the largest of which is the village of Archib (Chumakina, 2009b, p. 430). Additionally, unlike Avar and Lak, Archi does not have a long literary tradition. An orthography was developed by Moscow State University in 2006-2007 and accepted by the village, but “is not used by the Archis for any practical purposes” (Dobrushina, 2013, p. 380). While the Archi language is part of the Lezgian branch of the Northeast Caucasian family, Archis do not identify themselves as Lezgian,

and as Archi is a minority language, students are instead taught standard Avar in school, along with Russian (Chumakina, 2009b).

Historically, stable multilingualism was maintained in these villages as a result of endogamy, as described above, combined with close economic ties as a result of necessity, as described in Dobrushina (2013). The village of Archib is at a higher altitude than Chitab and Shalib, and therefore has poorer agricultural lands, but it apparently had better pastures, leading to economic exchange in which the Archis “rented fields and pastures to Chitab and Shalib, and bought crops from there, while selling them meat and wool” (Dobrushina, 2013, p. 383). The economic ties were close enough that those in the villages also participated in “guest-host relations” in which families in one villages would have a “partner family” in another village that they could stay with as needed, and in return they would host that partner family in their village for the purposes of trade and work (Dobrushina, 2013, p. 383). These close economic ties strongly encouraged multilingualism, while the strict practice of endogamy also stabilized that multilingualism and prevented large degrees of language shift.

Additionally, these three villages have historically fit the profile of asymmetrical multilingualism, as described in Nichols (2013). According to the Atlas of Multilingualism in Dagestan, for villagers born before 1919, 90% of those in Archib (native language: Archi) spoke Avar and 75% spoke Lak, but only 12% of those in Chitab (native language: Avar) spoke Archi and 0% of those in Shalib (native language: Lak) spoke Archi (Dobrushina, 2013). Archib is the economically poorer village and the more isolated, as both the Avar and Lak villages, Chitab and Shalib respectively, have linguistic and ethnic reinforcement from other villages in Dagestan. Additionally, Archi is looked down on by

those in Chitab and Shalib as a lower prestige language, with one person even referring to it as “the language of the devil” (Dobrushina, 2013, p. 387). As such, it is clear that language ideologies regarding the usefulness of speaking Archi have caused asymmetrical bilingualism between the Archib village and the Shalib and Chitab villages in which the Archi can speak Avar and Lak, the languages of Chitab and Shalib, but those in Chitab and Shalib do not or rarely learn Archi.

On the other hand, there has historically been fairly symmetrical multilingualism between the Shalib and Chitab villages. For speakers born before 1919, 60% of those in Chitab (native language: Avar) could speak Lak and 65% of those in Shalib (native language: Lak) could speak Avar (Dobrushina, 2013, p. 387). As mentioned earlier, both Avar and Lak are literary languages in Dagestan with historical power, and additionally they are less socially isolated, especially for the Avar in Chitab, who have other neighboring Avar-speaking villages. This lower social and ethnic isolation, as well as greater economic prestige, may have led to more equal social standing and prestige in the eyes of the speakers of Chitab and Shalib, leading to more stable, symmetrical bilingualism.

It is important to note that regardless of the symmetry or lack thereof of the multilingualism between Archi, Lak, and Avar, there is not one single language that acted as a local lingua franca for these villages (Dobrushina, 2013). It would have been possible for, for example, those in Shalib and Archib to speak Avar with those in Chitab, and those in Archib to speak Lak with those in Shalib, and have the Avar-speaking group in Chitab speak only Avar, the Shalib speak Avar and Lak, and the Archib to speak Avar, Lak, and Archi. However, this is not what was found in the Atlas of Multilingualism in Dagestan

(Dobrushina et al., 2017). Instead, what was discovered were stable patterns of symmetrical and asymmetrical multilingualism that did not involve a local lingua franca; villagers spoke the languages of their neighboring villages as needed on an economic basis, and communicated with each other in either of their primary languages, not in a third language.

### 2.1.3 Northeast Caucasian and the Comparative Method

Overall, the Northeast Caucasian language family exemplifies the challenges of the Comparative Method. Firstly, it is a large language family with dozens of languages; while not as large as the Austronesian language family mentioned earlier, there are still 40 or more languages to compare, depending on the selected splits between “dialect” and “language” of the researcher in question. Comparing a single reflex across every language would mean comparing approximately 40 words; comparing the number of reflexes necessary to adequately identify regular sound correspondences across the entire family would be in the thousands to tens of thousands of words. Additionally, individual languages within the family have experienced intense contact for thousands of years, not only from languages inside of the language family, but also from other languages outside of the family, such as Azerbaijani, Georgian, and Arabic, as well as Russian more recently. This intense contact, especially from languages within the same language family, can make it especially difficult to pick apart lexical borrowings from inherited vocabulary. Lastly, most of the Northeast Caucasian languages do not have historical written records; those that do, such as Avar and Dargwa, typically only go back several centuries and even then somewhat sparingly. While Udi has been recorded in the form of two palimpsests from 600 CE (Schulze, 2017) and Lak has a translated text from 924 (Friedman, 2020, p. 202),

these languages do not have continuous written records from eras until the present day; written records are still limited and scattered temporally. The overall majority of Northeast Caucasian languages do not have any written records, historical or otherwise, complicating the ability to understand and reconstruct these languages in their past forms.

As such, it makes sense that the Comparative Method has not been applied to the Northeast Caucasian language family in its entirety. For a thorough examination into the Comparative Method as applied to the languages of the Caucasus, see Schulze (2017); however, I will discuss a few relevant applications here. Kibrik and Kodzasov published two volumes in 1988 and 1990 gathering purported cognates from across the Northeast Caucasian language family with the first volume encompassing verbs and the second nouns (Kibrik & Kodzasov, 1988, 1990). However, no full etymological dictionary of the family has been published, nor has the proto-language been thoroughly reconstructed (Schulze, 2017). The book *A North Caucasian Etymological Dictionary* (Nikolayev & Starostin, 1994), which does attempt to create an etymological dictionary and reconstructions, combines the Northeast and Northwest language families into one family by creating cognate sets meant to maximize the similarities between the two families, creating incorrect Northeast Caucasian cognate sets in at least one-third of cases as a result, meaning they are not reliable etymologies (Nichols, 2003, p. 208). More work has been done to apply the Comparative Method in its entirety to specific sub-branches of the language family, but not to the family as a whole (Schulze, 2017). Nichols (2003) does utilize purported cognates from the Kibrik and Kodzasov volumes (1988, 1990) as well as Giginishvili (1977) to begin to identify the consonant sound correspondences in the Northeast Caucasian family, but focuses only on consonants and not vowels, and only does

preliminary and partial reconstructions of the vocabulary in proto-Northeast Caucasian. Additionally, the correspondences are based on only 50 items of vocabulary and a subset of the Northeast Caucasian languages (Nichols, 2003).

#### 2.1.4 Northeast Caucasian and Lexical Borrowing Studies

There have been some analyses of lexical borrowing in the Caucasus, and more specifically the Northeast Caucasian language family. These studies have frequently focused on either specific sub-branches of the language family or specific languages, or analyzed lexical borrowing from other language families into Northeast Caucasian. For example, Daniel et al (2021) analyzes loanwords from languages that have acted as local lingua francas in Dagestan and the Caucasus, such as Azerbaijani (Turkic), Avar (Northeast Caucasian), Georgian (South Caucasian), and Chechen (Northeast Caucasian), as well as loanwords from Russian, Persian, and Arabic, which did not have native speakers in the region but were or are languages with prestige, and loanwords from small-scale multilingualism of the type described above between Chitab, Shalib, and Archib. The findings indicate that lexical borrowings are more common from lingua francas than small-scale multilingualism, perhaps because the speakers may view the lingua francas as “linguistically neutral”, meaning usage of their lexical material would not be damaging to the ethnic or linguistic identity of the speakers (Daniel et al., 2021, p. 553). For example, a Lak and Avar speaker in communication with one another borrowing Russian lexical material could be viewed as more neutral than the Lak speaker borrowing Avar lexical material or the Avar speaker borrowing Lak lexical material. Overall, while this study does compare borrowing from small-scale multilingualism, Chitab, Shalib, and Archib are not among the villages examined, and the emphasis is on borrowing from larger lingua francas.

In another study, Chechuro et al. (2021) examines small-scale multilingualism, but focuses on loanwords from regional varieties of the Turkic language Azerbaijani into Northeast Caucasian languages and is therefore not an examination of lexical borrowing between Northeast Caucasian languages. A final study by Chechuro (2021) analyzes Russian, Azerbaijani, and Georgian loanwords into Rutul (Northeast Caucasian; Lezgian branch) speaking and Tsezic (Northeast Caucasian, Tsezic branch) speaking villages, emphasizing the importance of combining data on loanwords and multilingualism with historical information and cultural influence to describe outcomes of contact.

For borrowing into specific languages, the World Loanword Database does contain two entries from Northeast Caucasian Languages: Archi (Chumakina, 2009a) and Bezhta (Comrie & Khalilov, 2009). These entries contain approximately 1,400 items of vocabulary from each language along with a ranking from 1-5 of the likelihood of the item being borrowed, with 5 being ‘no evidence for borrowing’ and 1 being ‘clearly borrowed’ (Haspelmath & Tadmor, 2009). While this does give information about borrowing into Archi, and the accompanying chapter (Chumakina, 2009b) describes the semantic fields of the loanwords as well as how they are integrated into Archi morphologically and phonologically, the analyses is through the lens of Archi as a recipient language, not within a cluster of other languages.

Schulze (2013) discusses words for minerals and metals; as these words cannot be traced back to proto-Northeast Caucasian, the suggestion is that they were invented as needed in the already-diverged branches or borrowed from other language families. As such, while borrowing in this domain is discussed, the goal of the paper is not to analyze lexical borrowing per se.

In general, the Northeast Caucasian language family perfectly encompasses the difficulties of the Comparative Method: the large size of the language family, with over 40 languages; the lack of historical written records; and the intense language contact, with many sources for lexical borrowing, both internal to the language family and external. However, the Northeast Caucasian family is not the only family to have these difficulties. As such, many computational methods have begun to be applied to the problem of the Comparative Method, seeking to expedite some of the work involved.

## **2.2 Computational Methods for Identification of Cognates and Loanwords**

As can be seen from the discussion above, historical linguistics and the Comparative Method are complicated in their scope and application and, as a recursive and iterative process, the Comparative Method can also be time-consuming, further complicated by language contact and a lack of written records, both historical and present-day, for many languages. As a result, attempts have been made to automate the process.

Since the Comparative Method is not a single step process, instead comprising multiple unique and individual steps that each serve a separate function, computational methods for the Comparative Method have had to separately automate the different stages of the process. In general, the stages of the process that have been automated in various ways are automatic cognate detection, automatic borrowing detection, automatic reconstruction, and automatic family tree construction. For the purposes of this thesis, only automatic cognate and automatic borrowing detection will be discussed. For a more thorough overview of the broad history of applying computational methods to historical linguistics, I would recommend reviewing Joseph Rhyne's master's thesis titled



“Quantifying the Comparative Method: Applying Computational Approaches to the Balto-Slavic Question” (2017) and the preprint by Johann-Mattis List titled “Computational Approaches to Historical Language Comparison” to appear in the second edition of the Routledge Handbook of Historical Linguistics (List, preprint), as well as Language Classification by Numbers by April McMahon and Robert McMahon (2005).

### 2.2.1 Data Sources for Computational Methods

While some computational approaches take in large amounts of data from online dictionaries (e.g., St Arnaud et al., 2017; Steiner et al., 2011), in some cases with etymological information (e.g., Ciobanu & Dinu, 2020), many others focus the data on smaller, clearly defined lists of semantic concepts (e.g., Bergsma & Kondrak, 2007; Hall & Klein, 2011; Hantgan & List, 2018; Kassian, 2015; Rama, 2015). Often, these semantic concept lists are meant to represent “basic” vocabulary and can vary in length. The development of the Swadesh and then Leipzig-Jakarta lists, which are lists of “core” or “basic” vocabulary intended to be more resistant to borrowing (Tadmor et al., 2010), led the way for developing additional lists of “basic” vocabulary, often varying in size and tailored to a specific language family. The purpose of these lists can be multifold; they provide a base upon which to compare languages between and across language families more easily, and, if they truly contain data more resistant to borrowing, can be good candidates for the Comparative Method and the identification of regular sound correspondences.

## 2.2.2 Automatic Cognate Detection

Traditionally, as discussed previously, cognates have been determined manually by comparing words across languages and determining which of those express regular sound correspondences and are therefore candidates for being reflexes of the same ancestor word. In order to begin converting this process to a computational approach, and as discussed by Rhyne (2017) and List (preprint), many methods for detecting cognates involve the alignment of corresponding sounds between words as a first step in the process.

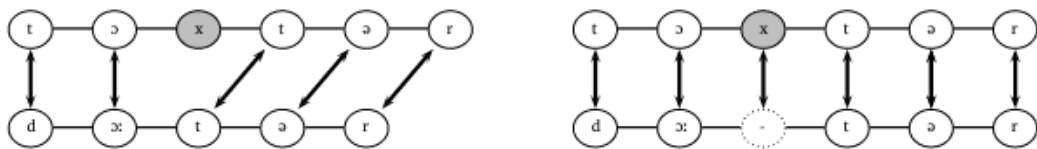


Figure 4: Alignment analysis of German *Tochter* and English *daughter* (List, 2012b)

As can be seen in Figure 4 from List (2012b), an alignment of German *Tochter* and English *daughter* involves identifying and aligning the segments that correspond with one another, inserting gap characters where necessary to account for deletions or insertions. In this example, there is a gap character inserted before the [t] in English *daughter* because this word does not have a character corresponding with the [x] in *Tochter*<sup>1</sup>.

For some cognate detection programs, the next step after alignment is to determine Levenstein, or edit, distances to determine the number of “steps” or “differences” from one

<sup>1</sup> The length marker, [:], is kept with the character it describes, [ɔ], for the purposes of alignment. While this segment could potentially be interpreted as [ɔɔ] with each [ɔ] being a separate character, the [t] would still not have a corresponding character and an additional gap would need to be included. For the purposes of this thesis, long consonants and vowels will be considered a single character and not two adjacent characters.

word to another. Taking the example of *Tochter* and *daughter* above, a transition from *Tochter* to *daughter* would require a substitution of [t] to [d], the lengthening of [ɔ] to [ɔ:], and the deletion of [x], meaning three changes. A change from *daughter* to *Tochter* would also be three changes, but three slightly different changes: substitution of [d] to [t], shortening of [ɔ:] to [ɔ], and inserting of [x].

Applying alignment and edit distances to cognate detection, Oakes (2000) utilized dynamic programming to identify the fewest number of operations (such as substitutions, insertions, and deletions) needed to change one form into another in bilingual word lists. The dynamic programming technique in Oakes's (2000) paper specifically identifies the optimal alignment from the word lists by finding the alignment that requires the least number of operations to change one word to another (Oakes, 2000). If the number of operations exceeded a set threshold of four, the pairs were considered not to be cognates. Issues with this approach appear to be the relatively arbitrary cutoff points for determining the number of operations needed to convert one cognate to another, as well as the reduced emphasis on identifying regular sound changes that make up the core of the Comparative Method. While the program can suggest regular sound correspondences based on identifying sound changes that occur more than a specific target number of times (the paper suggests two times as the cutoff), this again provides a relatively arbitrary cutoff and is in reality only identifying popular, or common, sound changes in the specific word list in question, not regular ones.

The ALINE program designed by Kondrak (2009) uses feature scores and saliency measures to compare the phonetic similarity of lexical items in bilingual word lists and establish cognates. A drawback of this method is that it can only take in bilingual wordlists

and not multilingual wordlists, so only two languages can be compared at a time (Kondrak, 2009), and when tested on the Balto-Slavic languages by Rhyne (2017), it under-detected cognates when compared to other systems tested in the same study.

The method described in Bouchard-Côté et al. (2013) was primarily designed for automatic reconstruction of proto-words, but it can also infer cognates using a context-dependent probabilistic string transducer<sup>2</sup>. Limitations include not being able to handle certain kinds of sound changes, such as metathesis, reduplications, and haplogogies, but the authors argue that as those changes tend to be less regular, they are also inherently less informative (Bouchard-Côté et al., 2013). When discussing Bouchard-Côté et al.'s (2013) approach, Atkinson (2013) does point out that the main purpose of the method is to compute automatic reconstructions; while the method can potentially infer the cognates from word string data, it does best when working with data from language families in which the genealogy is already fairly established, implying preexisting intensive utilization of the Comparative Method by historical linguists already.

In general, computational methods can struggle with identifying true cognates, meaning words that descended from ancestor words and not just words that are similar on the surface, as well as distinguishing cognates from borrowings.

#### 2.2.2.1 LingPy

While all computational methods for cognate and borrowing detection have specific uses or difficulties, one in particular that seems to attempt to recreate the steps of the

---

<sup>2</sup> A transducer is a model or set of rules that change a given input to an output. Transducers are able to account for many regular sound changes such as lenition and epenthesis, but are not able to account for more irregular sound changes, such as metathesis (Bouchard-Côté et al., 2013, p. 4225). A context-dependent probabilistic string transducer is one that can “encode a distribution over possible changes that a string might undergo as it changes through time” and that is “sensitive to the context in which a change takes place” (Bouchard-Côté et al., 2013, p. 4225).

Comparative Method in identifying regular sound correspondences and cognates and that is widely available for immediate implementation by linguists is the LexStat system created by Johann-Mattis List as a part of the LingPy library (List, 2012a; List & Forkel, 2021). The LingPy library is a python library for automating various tasks in historical linguistics that combines aspects of the Comparative Method with sequence comparison in order to automatically detect cognates and borrowings in wordlists (List, 2012a). While the LingPy library allows for a variety of tasks, two methods for cognate detection that have been implemented as a part of LingPy will be focused on here: Sound Class Alignment and LexStat (List, 2012a, 2012b).

The Sound-Class Based Phonetic Alignment (SCA) model builds off of work identifying sound classes by Dolgoplasky (1986) and was developed by List (2012b), based on the original idea that “phonetic correspondences inside a ‘type’ are more regular than those between different ‘types’” (Dolgopolsky, 1986 as cited in List, 2012b). For instance, Dolgoplasky proposed a type containing the labial obstruents [p], [b], and [f], with the implication being that changes from  $p < b$  or  $f < b$  would be more likely to occur than between [p], [b], [f], and other consonants (Dolgopolsky, 1986, p. 35). List (2012b) expands the sound classes from those present in Dolgopolsky (1986) and also includes vowels, which were initially omitted. The sound classes utilized for the SCA method can be seen in the table below:

Table 1: The SCA Sound Class Model (List, 2012b, p. 43)

No.	Cl.	Description	Examples	No.	Cl.	Description	Examples
1	A	unrounded back vowels	a, ɑ	15	P	labial plosives	p, b
2	B	labial fricatives	f, β	16	R	trills, taps, flaps	r
3	C	dental / alveolar affricates	ts, tʃ, tʃ, tʃ	17	S	sibilant fricatives	s, z, ʃ, ʒ
4	D	dental fricatives	θ, ð	18	T	dental / alveolar plosives	t, d
5	E	unrounded mid vowels	e, ε	19	U	rounded mid vowels	ɔ, o
6	G	velar and uvular fricatives	ɣ, x	20	W	labial approx. / fricative	v, w
7	H	laryngeals	h, ʔ	21	Y	rounded front vowels	u, ʊ, y
8	I	unrounded close vowels	i, ɪ	22	0	low even tones	11, 22
9	J	palatal approxoimant	j	23	1	rising tones	13, 35
10	K	velare and uvular plosives	k, g	24	2	falling tones	51, 53
11	L	lateral approximants	l	25	3	mid even tones	33
12	M	labial nasal	m	26	4	high even tones	44, 55
13	N	nasals	n, ŋ	27	5	short tones	1, 2
14	O	rounded back vowels	œ, ɔ	28	6	complex tones	214

However, while Dolgopolsky (1986) originally treated these ‘types’ as absolute in that transitions across types were not allowed, List (2012b) creates a scoring scheme that allows for transitions between types with weights. An example of the weighting of transitions between types can be seen in figure 5 below:

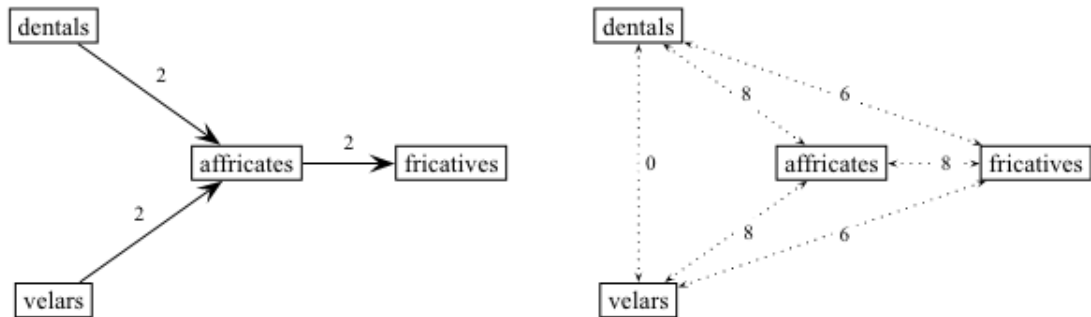


Figure 5: Modeling the directionality of sound change patterns in scoring schemes (List, 2012b, p. 40)

The similarity score for a segment compared to itself is set to 10; the similarity score for a transition from one class to another is then determined by subtracting the length of the shortest path between the two from the similarity score for a segment to itself (List, 2012b, p. 39). The paths between each class can be seen in the left portion of the diagram, and the

resulting similarity scores are in the right portion. For example, the similarity score of eight between velars and affricates is determined by subtracting the shortest path (two, as seen in the left portion) from ten, the similarity of a segment to itself. The similarity score of six between dentals and fricatives is determined by subtracting the length of its shortest path, four, from ten.

Overall, the basic workflow for the initial portion of the SCA method is: (1) tokenization, (2) class conversion, (3) alignment analysis, and (4) IPA conversion (List, 2012b, p. 41). An example of the workflow can be seen in figure 6 below.

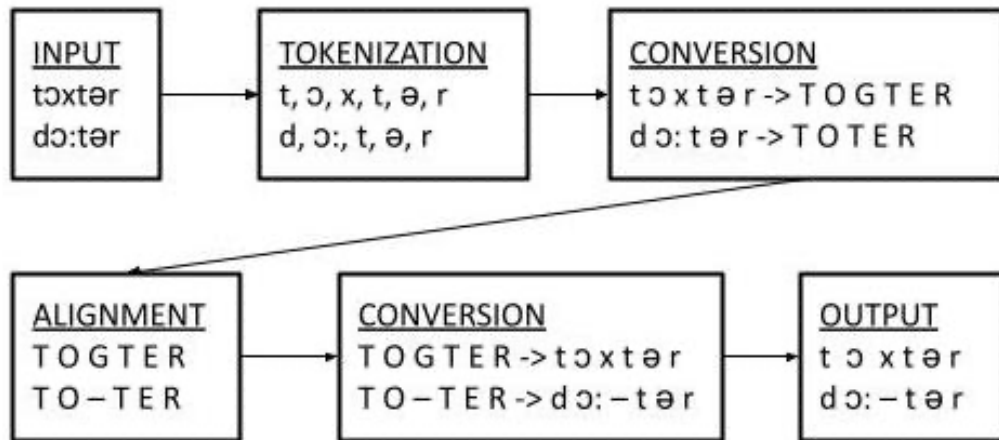


Figure 6: SCA Workflow [based on List (2012b, p. 42)]

In this example, the input words of [tɔxtər] and [dɔ:tər] ('daughter' in German and English, respectively) are first tokenized to separate each phoneme from the other. The phonemes are then converted into consonant classes, which are then aligned using SCA's alignment algorithm. Lastly, once the items are aligned, the sound classes are reconverted to their original IPA, maintaining the alignment with gap characters inserted where needed.

Once this sequence conversion takes place, the method then “takes the normalized Levenshtein distance between all word pairs in the same meaning slot and clusters these words into potential cognate sets using a flat version of the UPGMA algorithm, which terminates once a certain threshold of average distances between all words is reached” (J.-M. List et al., 2017, p. 4). The UPGMA originates from Sokal and Michener (1958) and utilizes a distance matrix between each item in the set, in this case being the “meaning slot” or semantic field (e.g., ‘mirror’ or ‘wife’), to identify the two items with the smallest distance between them. Those two items are then clustered together and a new distance matrix is calculated with distances between each item, including the new cluster as a single item. The algorithm can continue until every item in the meaning slot is clustered together, which may be useful for calculating phylogenetic trees; however, in this case one would not necessarily want every item to be clustered together, as not every item is necessarily genetically related. As such, a distance threshold is given such that the algorithm will stop clustering items once that threshold is reached. The specific threshold utilized is determined by the researcher, but testing by List et al (2017, p. 12) determined that the optimal threshold for the SCA method may be 0.45.

In comparison to the SCA method, the LexStat method utilizes four steps to “come close to the notion of sound correspondences in traditional historical linguistics” (List, 2012a, p. 120). The steps are as follows:

- (1) sequence conversion,
- (2) scoring scheme creation,
- (3) distance calculation, and
- (4) sequence clustering



(List, 2012a, p. 120)

The sequence conversion step utilizes the Sound-Class-Based Phonetic Alignment developed by List (List, 2012b), as with the SCA method described above. However, the difference between the methods is that in the later steps the LexStat method uses language-specific scoring schemes “to derive a distribution of sound-correspondence frequencies under the assumption that both languages are not related” (J.-M. List et al., 2017, p. 5). After the language-specific scoring schemes are created, the distances between word pairs are calculated (List, 2012a, p. 122). An example of distance calculations for German-English word pairs using the SCA and LexStat methods can be seen in figure 7 below, where a higher number indicates a greater distance between the two forms:

<b>Word Pair</b>			<b>SCA</b>	<b>LexStat</b>
German	<i>Schlange</i>	[ʃlanɣə]	0.44	0.67
English	<i>Snake</i>	[sneɪk]		
German	<i>Wald</i>	[valt]	0.40	0.64
English	<i>wood</i>	[wud]		
German	<i>Staub</i>	[ʃtaup]	0.43	0.78
English	<i>dust</i>	[dʌst]		

Figure 7: SCA distance versus LexStat distance (List, 2012a, p. 122)

As described in List (2012a, p. 122), these distance calculations show the benefits of LexStat; while the scores created utilizing the SCA method are similar and fairly low for all word pairs due to the surface similarity they contain, the LexStat distances are notably larger. As none of these items are genetically cognate, these results imply that the language-specific scoring scheme of the LexStat method may be better able to determine

the difference between true cognates and lexical items that merely demonstrate surface similarity (List, 2012a, p. 122). Once the distances between items have been calculated, the same UPGMA algorithm used for the SCA method clusters the items into cognate sets based on a researcher-provided threshold (List, 2012a, p. 122). Research by List et al. (2017, p. 9) shows that the ideal threshold for the LexStat method may be .60.

### 2.2.3 Automatic Borrowing Detection

The investigation of computational borrowing detection has been more limited than the investigations of cognate detection (List & Forkel, 2022, p. 3). Ciobanu & Dinu (2015) describe a methodology devised to discriminate between borrowings and cognates by first aligning the words, then extracting orthographic features from the aligned words and finally determining whether the pairs are cognates or borrowings through training a binary classifier. While they were able to correctly determine cognates to a fairly high degree for three out of the four pairs of languages they investigated, the model relied on orthographic cues in making its decision, meaning it may be useful only for written languages. Essentially, the model took in written words instead of IPA transcriptions, as many other approaches do, allowing the orthography to play a large role in determining the differences between the loanwords and cognates. Additionally, although they define a cognate as a pair of words that “share a common ancestor and have the same meaning” (Ciobanu & Dinu, 2015, p. 431), this definition does not take into account semantic change, which is frequent across languages. Many times, items that are true cognates can actually appear very different and have dramatically dissimilar meanings, while items that are phonetically and semantically similar are more expected to be the result of borrowing or chance (Kondrak, 2009, p. 203).

The difficulty in distinguishing loanwords from inherited vocabulary in genetically related languages is one of the primary difficulties of automating borrowing detection and a big part of the reason why it remains “unsolved” according to Jäger (2019, p. 178). Lexical items borrowed from related languages may appear as cognates when using automatic cognate detection methods because the algorithms are unable to accurately distinguish between words that are similar because they are the result of an unbroken chain of inheritance or because they are borrowed. Additionally, as discussed in List (2019, p. 13), most computational methods for detecting borrowing and language contact focus on phylogeny and “distribution-related conflicts and borrowability”, but less on words that, for example, violate known sound correspondences. As such, more effort to combine the cognate detection methods and borrowing detection methods could be useful to improve these outcomes.

In some ways, automatic borrowing detection can be seen as the other side of the coin from automatic cognate detection. While a cognate detector identifies some words as cognate, there is a question of what to do with the words that are not cognate. Those words that are not cognate may be either the result of semantic drift (meaning a word in language B may be cognate with some other word in language A that is not the word it is being compared with) or borrowing from another language. The Comparative Method itself doesn't focus on identifying borrowings, but doing so is an inevitable side effect of the method. When identifying words with regular sound correspondences, there are going to be words that do not fit the pattern, and those must be reconciled with the history of the language involved. As such, automatic borrowing detection and cognate detection can be

seen as going hand in hand; similar methods may be used, but it is the end result and final goal that is different.

Another aspect of the challenge in automating borrowing detection is the difference in traditional methods used to identify cognates versus borrowings. Evidence in favor of the cognacy of two words can be seen in the realization of regular sound correspondences, which, while still challenging to implement in an automated manner, is fairly straightforward to understand and is applicable cross linguistically in a binary manner. However, the evidence required to determine that a lexical item has been borrowed is far more nuanced and usually relies on the cumulative combination of lots of smaller pieces of evidence, such as cultural context, type of word borrowed, the phonotactics of the languages involved, and more, which is far less suitable to being automatized (List, 2019, p. 13).

#### 2.2.4 Automated versus Computer-Assisted Approaches

While the presence and creation of computational methods for historical linguistics has dramatically increased over the past two decades (List, 2019, p. 1), it is the understanding of many that these computational approaches cannot entirely replace the abilities and knowledge of trained linguists (Rhyne, 2017, p. 92; Steiner et al., 2011, p. 122). As such, some are arguing for a “computer-assisted” approach to historical linguistics (List et al., 2017; Wu et al., 2020). This approach can function in a variety of ways depending on the assistance desired or utilized by automated computational methods. For instance, List (2017) suggests utilizing a cognate detecting algorithm as a first-pass over language data and then manually verifying and correcting the cognate results

afterwards before moving to later steps, such as reconstructing a language family tree. Wu et al (2020) lay out a pipeline in which raw data is tokenized, cognates are identified within the same semantic slots, those cognate sets are phonetically aligned, cognates are compared across semantic meanings, and then sound correspondences are identified; while each step utilizes computational methods, the data can be checked and, as necessary, corrected by experts between each step.

The above two computer-assisted approaches seek to identify cognates; however, given the particular difficulties in automating borrowing detection, this may be the area in which a computer-assisted approach could provide the greatest benefit. In one such example, Moro et al (2023) utilize both computational and qualitative methods to investigate borrowings from Papuan languages into Alorese, an Austronesian language. The authors first run their data through LexStat (List, 2012a), the previously described cognate detection program, stating that while LexStat was designed to detect cognates, it in general does a good job of detecting similarities between lexical items, allowing it to identify potential loanwords as well (Moro et al., 2023). Then, the results of the LexStat program are filtered according to pre-decided hypotheses regarding the potential lexical distribution of loanwords; for example, if a cognate class contains lexical items present in Alor-Pantar (AP) languages and absent in Alorese, Indonesian, Flores-Lembata (FL), and other Austronesian languages, then it is likely inherited vocabulary in those Alor-Pantar languages (Moro et al., 2023, p. 220). However, if a cognate class contains lexical items present in Alor-Pantar languages and Alorese but absent in Indonesian, Flores-Lembata, and other Austronesian languages, then this is likely evidence of a loan word from an Alor-Pantar language into Alorese or vice versa (Moro et al., 2023,

p. 220). The full table of hypotheses based on lexical distribution in Moro et al. (2023, p. 220) can be seen below:

Table 2: “Relevant patterns of distribution of lexically similar forms in languages of the region, and the corresponding borrowing or inheritance history of such a form” (Moro et al., 2023, p. 220)

Hypothesis	AP languages	Alorese	Flores-Lembata	Indonesian	Other Austronesian	Explanation
1	Present	Absent	Absent	Absent	Absent	Inherited TAP vocabulary
2	Present	Present	Absent	Absent	Absent	Loan from AP into Alorese or vice versa, to be further inspected
3	Present	Present	(likely present)	Present	(likely present)	Indonesian loan into local languages
4	Present	Present	Present	Absent	Absent	Likely Alorese loan (inherited from PFL) into AP

By using LexStat to identify the lexical similarities between words and then filtering the results based on the expected lexical distribution of the items that the authors want to study, in this case loanwords from AP languages into Alorese, they are able to spend less time on this initial portion of the process and instead focus their time and efforts on the analysis of the results. Such a computer-assisted approach can save time and energy by bringing to the forefront the lexical items that will be the most interesting and useful to the researchers, allowing them to focus on this smaller list of items in applying the more qualitative methods for borrowing detection. Overall, computer-assisted approaches can save researchers time, especially when large language families and datasets are involved, and may be more immediately useful in the present when compared to fully automated approaches, as many fully automated approaches still require further testing.

### 2.2.5 Computation Methods Applied to the Caucasus

Some attempts to apply computational methods for historical linguistics to the Caucasus have been made; however, to my knowledge, these methods have only been applied to specific sub-branches of the Northeast Caucasian family, and not to the family as a whole. For example, Kassian (2015) tests phylogenetic methods on the languages of the Lezgian branch utilizing cognate lists compiled in two ways: through the traditional Comparative Method, and through consonant classes. The consonant classes method utilizes the consonant classes developed by the *Global Lexicostatistical Database* project (Starostin, 2011) and marks forms as cognate with each other automatically if their consonant class transcriptions match, although he also notes that this is not “true” cognate detection, as it could lead to words not being marked cognate even if they do in actuality descend from the same root (A. Kassian, 2015, p. 5). A later paper by Kassian (2017) also tests various phylogenetic methods, this time on the Lezgian and Tsezic branches, with preformed cognate lists. Lastly, Zaitsev and Minchenko (2022) utilize a logistic regression model to computationally detect borrowings in eight dictionaries of various Andic languages. Overall, while computational methods have been applied to the Caucasus, the scope has been rather limited in terms of methods applied as well as number of languages compared.

## CHAPTER 3. DATA AND METHODS OF THE PRESENT STUDY

The goal of this thesis is to apply a computer-assisted approach to investigate lexical borrowing in the languages of the Caucasus, and in particular between the languages spoken in the villages of Chitab, Shalib, and Archib (Avar, Lak, and Archi, respectively). Utilizing data from DagSwadesh (Filatov & Daniel, n.d.) and the Intercontinental Dictionary Series (2023) in connection with computational methods for cognate detection designed by List (2012a, 2012b), as well as the lexical distribution approach for investigating loanwords from Moro et al (2023), I seek to identify possible loanwords and investigate them qualitatively without having to individually examine and compare the thousands of lexical items present in the Intercontinental Dictionary Series (2023).

### 3.1 Data

One of the critical aspects of computational methods for linguistics is the data that goes into these systems. Having quality, consistent data that is reliable is critical for the Comparative Method in general, but may perhaps be even more critical for computational approaches to historical linguistics, as the systems themselves can only do exactly what they are programmed to do and are not able to evaluate the accuracy of the data they are fed. If a historical linguist is working with poor or inconsistent data they may be able to recognize it part way through, but an algorithm does not necessarily have such a function. The data sources utilized for the lexical information in this project are DagSwadesh (Filatov & Daniel, n.d.) and the Intercontinental Dictionary Series (2023). The Atlas of Multilingualism in Dagestan (Dobrushina et al., 2017) is also utilized for information regarding patterns of multilingualism in specific villages in Dagestan.



### 3.1.1 DagSwadesh

DagSwadesh (Filatov & Daniel, n.d.) is one of the datasets utilized for this project. The database includes 110-word Swadesh lists for 21 lects from the Avar-Andic branch of the Nakh-Daghestanian language family in Dagestan that were collected on the village level (Filatov & Daniel, n.d.). In this case, “lect” is used to encompass what may traditionally be seen as separate languages or dialects of a language. As stated on their website, this database fills “gaps in the existing datasets” of Northeast Caucasian languages by working on the village level because it shows possible “differences between villages speaking what is conventionally seen as one and the same language” (Filatov & Daniel, n.d.). Having the list be on the village level means it will likely be more accurate to actual language use in the community, as opposed to a dictionary or grammar, which could maintain more prescriptivist or conservative accounts of language use, or which may be of a different language variety than the specific villages in question.

The original Swadesh List was developed in 1952 by Morris Swadesh who, seeking to compare the relatedness of languages by identifying a constant rate at which the vocabulary of a language changes and is replaced, created a list of 215 lexical items that he considered stable and less likely to change (1952, p. 455). He later refined the list several times in order to attempt to make it as universal and “culture-free” as possible, eventually ending with a list of 100 concepts (Tadmor et al., 2010, p. 228). Other lists of basic vocabulary have also been created throughout the years, perhaps most notably the Leipzig-Jakarta list, which was created through cross-linguistic investigation of which concepts are actually most resistant to borrowing, resulting in a 100 word list with some overlap with the original Swadesh list (Tadmor et al., 2010). Overall, the lexicostatistical

approach and glottochronology, which came after the advent of the Swadesh list, have been criticized both due to the difficult-to-define nature of a “basic vocabulary” as well as the erroneous assumption that lexical items change at a consistent rate across languages<sup>3</sup>. However, having a list of concepts that are seen as at least less susceptible to borrowing, if not impervious, has still been useful as a way to compare vocabulary across languages in a consistent manner, and these lists have also functioned as a starting point for quantitative approaches (Campbell, 2013).

The DagSwadesh dataset contains lexical data with cognate identifications already assigned for the 100-item Swadesh list (Filatov & Daniel, n.d.). Since this dataset includes gold-standard, expert assigned cognate information, it is ideal for testing LingPy’s cognate detection methods and thresholds. The method and threshold that achieves the best results for this data (discussed in section 2.1 below) is then applied to a larger dataset of Northeast Caucasian lexical information, the Intercontinental Dictionary Series (2023).

### 3.1.2 Intercontinental Dictionary Series

The additional lexical data for this project comes from the Intercontinental Dictionary Series, a database created by Key & Comrie for the purpose of organizing lexical material such that cross-linguistic comparisons can more easily be made (2023). The database uses a 1,310-item concept list arranged by topic and contains dictionaries for 82 different lects of Northeast Caucasian, covering what have traditionally been seen as different languages as well as various dialects (*The Intercontinental*

---

<sup>3</sup> The rate at which languages change and replace lexical material is not constant; for example, intense contact can speed up lexical replacement, while isolation can slow it down (Heggarty, 2010, p. 304). Many geographic, demographic, social, political, and cultural factors can affect the replacement of lexical material, and as these factors themselves are not stable over time, it would be erroneous to expect the rate of lexical replacement to be as well (Heggarty, 2010, p. 304).

*Dictionary Series*, 2023). The actual number of lexical items in each dictionary for the Northeast Caucasian lects varies from approximately 1,300 to 1,700 items, as many concepts contain more than one lexical item to encode synonyms and some concepts are left blank (*The Intercontinental Dictionary Series*, 2023). Critically, the Intercontinental Dictionary Series contains dictionaries for Archi (Khalilov, 2023a, 2023b) and the lect of Lak spoken in Shalib (Khalilov, 2023e), as well as standard Avar (Khalilov, 2023c) and standard Lak (Khalilov, 2023d), allowing it to serve as the starting point of investigating lexical borrowing in Archib, Shalib, and Chitab (although it unfortunately does not contain the specific variety of Avar spoken in Chitab). It also contains dictionaries of Kumyk, Azerbaijani, Persian, and Russian, frequent external sources of lexical borrowing into the Northeast Caucasian family.

### 3.1.3 Atlas of Multilingualism in Dagestan

Utilizing the method of retrospective family interviews, Dobrushina (2013) is able to reconstruct patterns of multilingualism in Dagestan up to 150 years ago. The data from these interviews are part of the Atlas of Multilingualism in Dagestan and contain information about the languages spoken by individuals, their parents, and grandparents in clusters of villages (Dobrushina et al., 2020a). As Dagestan is a region that has undergone rapid shift from small-scale, neighbor multilingualism to having Russian as a lingua franca (Dobrushina & Kultepina, 2021), the use of these interviews is critical in describing and preserving knowledge of the patterns of multilingualism in the area. The project has also investigated the patterns of multilingualism specifically in Archib, Shalib, and Chitab (Dobrushina et al., 2020a), which is why those villages were selected for this project. As scale, intensity, and symmetry of multilingualism is one factor that can affect the outcomes

of language contact, including the prevalence of lexical borrowing (Thomason, 2001), having this data available is useful in interpreting possible instances of lexical borrowing within the present study.

### **3.2 Computational Methods**

The next step in this computer-assisted approach is to utilize computational methods as a first-pass over the data. The Intercontinental Dictionary Series dictionaries for the lects of investigation in this project (Lak Shali, standard Lak, standard Avar, and Archi) contain a total of 7,358 lexical items. Lak Shali is the variety, or lect, of Lak spoken in Shalib, standard Lak is the standard variety, standard Avar is the standard variety of Avar, and Archi is the lect spoken in Archib. As mentioned earlier, the dialect of Avar spoken in Chitab is not a part of the Intercontinental Dictionary Series. With the dictionaries for Russian, Kumyk, Azerbaijani, and Persian added in order to filter out borrowings from external sources, that brings the total number of lexical items to 14,334. Borrowings from these external sources are filtered out in order to keep the focus on internal borrowings between Avar, Lak, and Archi specifically. Arabic is another frequent external source of loanwords, but could not be added to the data to be filtered out as the Intercontinental Dictionary Series does not contain a dictionary for Arabic. As a result, any loans from Arabic will need to be filtered out at the later investigative stage. A computer-assisted approach to filtering out external borrowing and identifying possible internal borrowings allows me to avoid having to individually analyze every item; instead, the most relevant items are brought to the forefront for that closer, qualitative investigation.

### 3.2.1 Evaluating Cognate Detection Methods

In order to evaluate the accuracy of the LexStat and SCA methods, as well as identify an appropriate threshold for cognate clustering, I apply both methods for cognate detection from the LingPy library to the DagSwadesh dataset across a range of thresholds: 0.40 to 0.65 in intervals of .05 for the SCA method, and 0.55 to 0.80 in intervals of .05 for the LexStat method. The output of each method at each threshold can then be evaluated for precision, recall, and f-score (or combined score). Precision refers to how many of the returned cognates are true cognates. For instance, if a particular method at a particular threshold returns 100 items as cognates but only 80 of them are actual cognates, then the precision would be .80 or 80%. Recall refers to how many of the true cognates are actually returned. For instance, if a particular method at a particular threshold returns 80 true cognates and leaves 20 of them behind, then the recall would be .80 or 80%. Essentially, precision and recall determine the extent of false positives, false negatives, true positives, and true negatives.

For the purpose of calculating precision and recall, the LingPy library utilizes B-Cubed scores (List et al., 2017, p. 7). B-Cubed scores calculate the precision and recall for each individual item and then average those precision and recall numbers across all items in order to determine the overall precision and recall (Amigó et al., 2009, pp. 471–472). An example of how precision and recall are computed for a single item can be seen in figure 8 below:

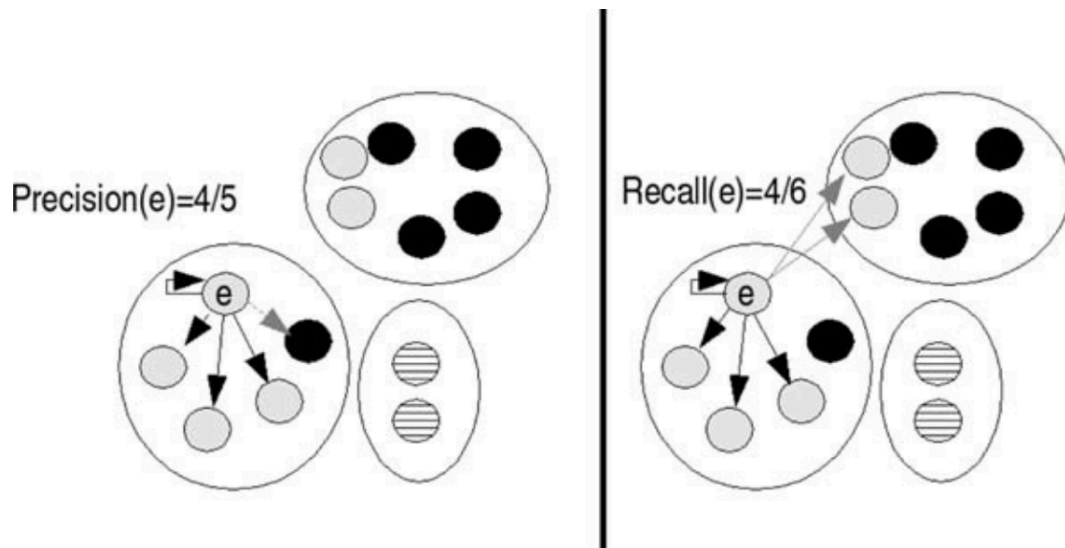


Figure 8: Example of computing the B-Cubed precision and recall for one item (Amigó et al., 2009, p. 471)

In this figure,  $e$  is the item being evaluated. On the precision side, only the cluster containing  $e$  is examined. Within this cluster, four out of the five items are the same category as  $e$  (including itself), so the precision for  $e$  is  $4/5$  or .80. On the recall side, only items in the same category as  $e$  are examined. Within this category, four out of the six items of the same category as  $e$  are clustered with  $e$  (including itself), so the recall is  $4/6$  or .66. Once the precision and recall have been evaluated for each individual item, the overall B-Cubed precision is calculated by averaging the precision scores for each item, and the overall B-Cubed recall is calculated by averaging the recall scores for each item (Amigó et al., 2009, p. 472).

Having a precision and recall score for each method at each threshold to be tested is useful for identifying the extent of false positives and false negatives. These two scores can also be combined into one overall score to encompass both precision and recall, known

as an F-Score (List et al., 2017, p. 8). The F-Score can be calculated as follows, with  $P$  referring to precision and  $R$  referring to recall (List et al., 2017, p. 8):

$$F = 2 \times \frac{P \times R}{P + R}$$

The LingPy python library contains a script that can take the output from its cognate detection program and a gold standard cognate list for the same data and compute the B-Cubed precision and recall scores as well as the F-Score, making the evaluation of its methods for data in which expert cognate decisions have already been made straightforward (List & Forkel, 2021).

Overall, The LingPy SCA and LexStat methods are both individually applied to the DagSwadesh data, which has expert cognate identifications, across various cognate thresholds in order to evaluate which method and threshold combination is most effective for that data. Then, that method and threshold combination can be applied to the Intercontinental Dictionary Series data, with a focus on lects of Avar, Lak and Archi spoken in the clusters of villages surveyed in the Atlas of Multilingualism in Dagestan (Dobrushina et al., 2020a).

### 3.2.2 Applying the Selected Method to the Village Cluster

Once the optimal method and threshold is identified from testing on the DagSwadesh (Filatov & Daniel, n.d.) dataset, that method and threshold is applied to the data from standard Avar, standard Lak, Lak Shali, and Archi, the languages of investigation, as well as data from Russian, Persian, Azerbaijani, and Kumyk, common sources of external borrowing, in the Intercontinental Dictionary Series (2023). The lexical

distribution of the output is then investigated to highlight and bring to the forefront those items that are most likely to be loanwords.

The following table gives an overview of the lexical distribution patterns under examination across the languages involved, where Non-NEC stands for “non-Northeast Caucasian”, representing the external sources of loanwords mentioned above<sup>4</sup>.

Table 3: Possible distribution patterns of lexical items and hypothesized explanations

Option	Archi	Avar	Lak	Lak Shali	Non-NEC	Hypothesis
1	Present	Present	Absent	Present	Absent	Loanword from Avar into Archi and Lak Shali
2	Present	Present	Absent	Absent	Absent	Loanword from Avar into Archi
3	Absent	Present	Absent	Present	Absent	Loanword from Avar into Shali dialect of Lak
4	Present	Absent	Present	Present	Absent	Loanword from Lak into Archi
5	Present	Absent	Absent	Present	Absent	Loanword from Archi into Shali dialect of Lak

As indicated in row one, If a lexical item appears in Archi, Avar, and Shali, but not in standard Lak, then the word is hypothesized to be a loanword from Avar into both Archi and Shali. As indicated in row two, items that appear solely in Avar and Archi are hypothesized to potentially be loanwords from Avar in Archi. As indicated in row three, items that appear in Avar and only the Lak Shali data are hypothesized to be potential

<sup>4</sup> While it is possible for words of Russian, Persian, Arabic, or Turkic origin to have entered into, for example, Archi through contact with Avar instead of directly from the source languages, I follow Chechuro et al. (2021) and Daniel et al. (2021) in removing these items from analysis, as determining the path of borrowing for these external items can be exceedingly difficult.



loanwords from Avar into Lak Shali. If a lexical item appears in Archi and both standard and Shali dialects of Lak, but not in standard Avar, then the lexical item is hypothesized to be a loanword from standard Lak into Archi, as indicated in row four. Lastly, In order for a lexical item to be identified as a possible loanword from Archi into the Shali dialect of Lak, it must appear in Archi and Shali, but not in Avar or standard Lak. While the item could theoretically also be a loan from Lak Shali into Archi, I would then expect it to also appear in standard Lak. For each row, the item must also be absent from the potential sources of external borrowing.

There are, of course, other possibilities, such as a lexical item that appears in only standard Avar or standard Lak, or in all four lects. Items that appear in Archi, Avar, standard Lak and Shali were omitted because there would be too many hypotheses: the word could be a loanword from Avar into Lak and Archi or from Lak into Avar and Archi, a word inherited from proto-Northeast Caucasian, or a loanword into all three languages. There is also the theoretical possibility that the word could be a loanword from Archi into both Lak and Avar, but this is highly unlikely, as Avar and Lak are both more prestigious languages spoken by far higher numbers of people across much larger geographic areas. Regardless, the focus is on the combinations suggesting loanwords from Avar into the Shali dialect of Lak and Archi, Lak into Archi, and from Archi into the Shali dialect of Lak. The combinations listed above are the most useful for determining lexical effects of contact in the languages spoken in these three villages without access to a Chitab Avar-specific lexical list.

## CHAPTER 4. RESULTS

In this chapter, I will review the results of testing the LexStat and SCA methods on the DagSwadesh dataset as well as discuss the outcomes of each hypothesis regarding the lexical distribution of family-internal loanwords and the implications for a computer-assisted approach to lexical borrowing detection.

### 4.1 Testing on DagSwadesh

The first step is to test LingPy's cognate detection software on the DagSwadesh dataset, as it already contains expert cognate identification (Filatov & Daniel, n.d.). Tables 4 and 5 below contain the B-Cubed precision and recall as well as the F-Scores for the SCA method and LexStat method respectively across different cognate clustering thresholds.

Table 4: SCA Method Precision, Recall, and F-Scores

Threshold	Precision	Recall	F-Score
.40	0.9245	0.7866	0.8500
.45	0.8950	.8481	.8709
.50	.8670	.9065	.8863
.55	.8330	.9424	.8849
.60	.7939	.9644	.8709
.65	.7474	.9759	.8465

Table 5: LexStat Method Precision, Recall, and F-Scores

Threshold	Precision	Recall	F-Scores
.55	.9639	.7570	.8480
.60	.9404	.7917	.8597
.65	.9204	.8422	.8796
.70	.8899	.9045	.8972
.75	.8479	.9404	.8918
.80	.8029	.9574	.8734

The data in these tables can also be represented as the following figures:

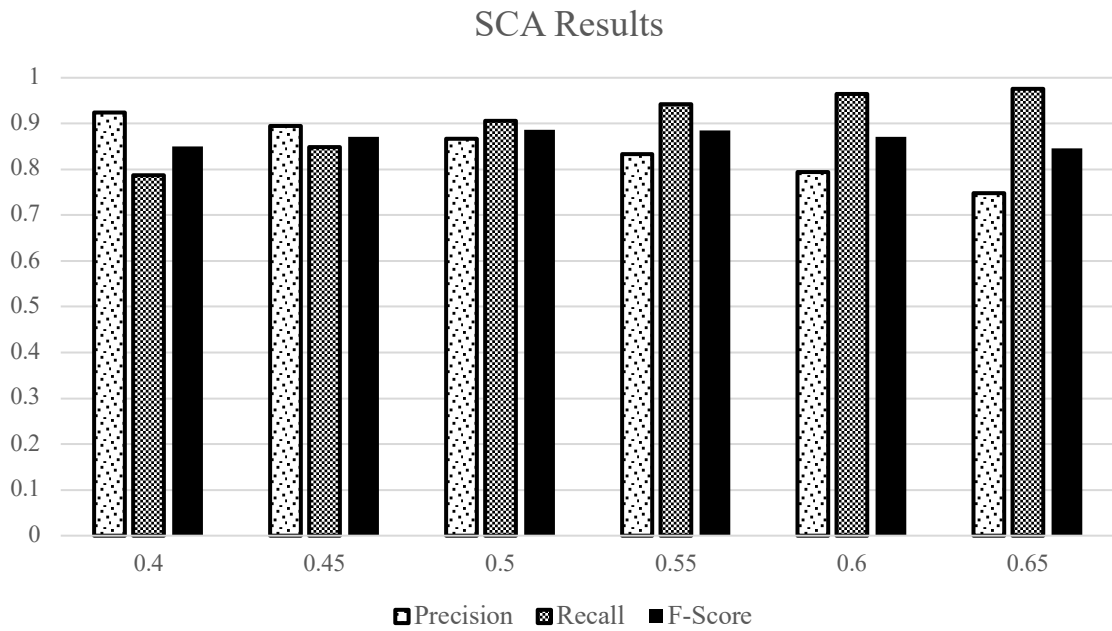


Figure 9: SCA Results

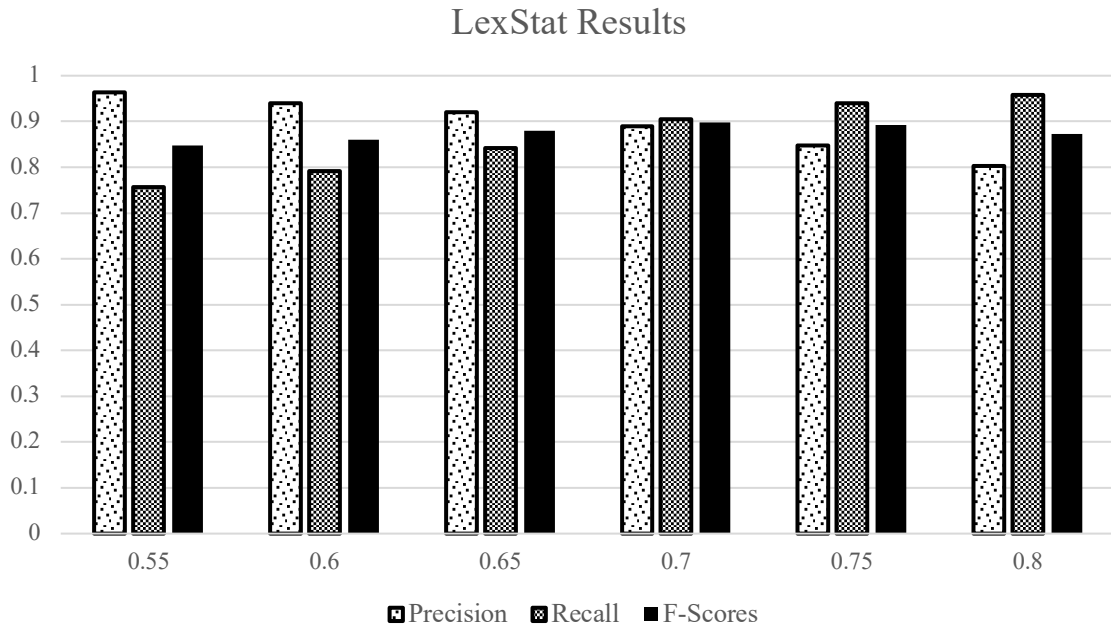


Figure 10: LexStat Results

As can be seen from the tables and charts above, precision and recall are inversely related. As discussed in section 2.2.2.1, the threshold is the distance at which the UPGMA clustering algorithm will cease clustering. Therefore, a lower threshold will only group together the most similar items, while a higher threshold will allow items that have a greater distance between them to be clustered together. With lower thresholds, precision is higher, meaning cognates that are grouped together are more likely to actually be cognates according to the expert identifications; however, the recall is also low at these same thresholds, meaning some true cognates are being left out of the clusters. As the threshold increases, precision drops and recall improves. By looking at the combined scores, or f-scores, it is possible to determine what threshold allows for the most ideal balance of precision and recall. Alternatively, a researcher could decide that precision or recall are

more important to the question being answered and choose to use a threshold that emphasizes that score.

#### **4.2 Implementation on the Intercontinental Dictionary Series**

Based on the results in the previous section as well as the previous study in Moro et al. (2023), I elected to perform the LexStat method on the Intercontinental Dictionary Series (2023) data. The threshold of .70 was initially selected due to having the highest f-score across all methods and thresholds. However, after reviewing the output of the files at a threshold of .70, there were many erroneous clusters. As such, I elected to use the LexStat approach at a threshold of .60 as recommended by List et al (2017) and utilized by Moro et al (2023). Lowering the threshold may negatively impact the recall such that not all similar sets would be collected, but it will also positively affect the precision by reducing the number of false positives. The Intercontinental Dictionary Series data from standard Avar, standard Lak, Lak Shali, and Archi are used as the villages focused on are Archib, Shalib, and Chitab. The database contains two varieties of Archi without explanation; this is unusual, as Archi is a fairly small language in terms of number of speakers (approximately 1,200) and is spoken in six to seven settlements, including Archib, that are all within walking distance of one another (Chumakina, 2009b). I have elected to use both Archi varieties and treat them as “Archi” as a whole, as there is significant overlap in the lists. Unfortunately the database does not contain lexical material for the dialect of Avar spoken in Chitab; as such, the lexical material from this database can be used to identify borrowings from standard Avar into dialectal Lak spoken in Shalib and into Archi and borrowing between the dialectal Lak spoken in Shalib and Archi, but not borrowings from the dialectal Lak spoken in Shalib and Archi into the dialectal Avar spoken in Chitab.

To reiterate the hypotheses based on lexical distribution discussed in section 3.2.2, in order for a lexical item to be identified as a possible loanword from Archi into the Shali dialect of Lak, it must appear in Archi and Shali, but not in Avar or standard Lak. If a lexical item appears in Archi and both standard and Shali dialects of Lak, but not in standard Avar, then the lexical item is hypothesized to be a loanword from standard Lak into Archi. If a lexical item appears in Archi, Avar, and Shali, but not in standard Lak, then the word is hypothesized to be a loanword from Avar into both Archi and Shali. These hypotheses are articulated in Table 1 below. There are, of course, other possible combinations, such as items that appear only in standard Avar and standard Lak or that appear in all four, and other explanations, such as loanwords from a different language. However, items that appear only in standard Avar and standard Lak would not be useful for analyzing lexical contact on the village level. Items that appear in Archi, Avar, standard Lak and Shali were omitted because there would be too many hypotheses: the word could be a loanword from Avar into Lak and Archi or from Lak into Avar and Archi, a word inherited from proto-Northeast Caucasian, or a loanword into all three languages. There is also the theoretical possibility that the word could be a loanword from Archi into both Lak and Avar, but this is highly unlikely, as Avar and Lak are both more prestigious languages spoken by far higher numbers of people across much larger geographic areas. Regardless, the focus is on the combinations suggesting loanwords from Avar into the Shali dialect of Lak and Archi, Lak into Archi, and from Archi into the Shali dialect of Lak. The label “Non-NEC” in the table represents Non-Northeast Caucasian sources of potential borrowing: Russian, Persian, Kumyk, and Azerbaijani. Arabic could not be included despite its status as a source of borrowings in Northeast Caucasian

languages because it does not at this time have a dictionary in the Intercontinental Dictionary Series (2023). The combinations listed below are the most useful for determining lexical effects of contact in the languages spoken in these three villages without access to a Chitab Avar-specific lexical list.

Table 6: Possible distribution patterns of lexical items and hypothesized explanations

Option	Archi	Avar	Lak	Lak Shali	Non-NEC	Hypothesis
1	Present	Present	Absent	Present	Absent	Loanword from Avar into Archi and Lak Shali
2	Present	Present	Absent	Absent	Absent	Loanword from Avar into Archi
3	Absent	Present	Absent	Present	Absent	Loanword from Avar into Shali dialect of Lak
4	Present	Absent	Present	Present	Absent	Loanword from Lak into Archi
5	Present	Absent	Absent	Present	Absent	Loanword from Archi into Shali dialect of Lak

Out of 1,301 total concepts (the semantic values attached to the lexical items investigated, such as ‘storm’ and ‘again’) and 7,358 lexical items distributed across the four (five, if counting Archi 1 and Archi 2 separately) lects, the number of words, and by extension possible loanwords, identified matching the hypotheses varies widely across hypotheses. For hypothesis one (row one in table six above), there are only 32 lexical items detected as similar across eight concepts, meaning there are 32 similar words in Avar, Archi and Shali that do not appear in the standard Lak list. Hypothesis two (row two in table six above), which would possibly demonstrate loanwords from Avar into Archi, contains 396 items across 140 concepts. Hypothesis three (row three in table six above),

which would possibly demonstrate loanwords from Avar into Shali, contains 29 items across thirteen concepts. Hypothesis four (row four in table 6 above), which would possibly demonstrate loanwords from Lak into Archi, contains 228 items across 59 concepts. Lastly, hypothesis five (row five in table 6 above), which would possibly demonstrate loanwords from Archi into Shali, contains 31 items across twelve concepts.

However, these counts are only the initial overview. As stated earlier, these computational methods are best utilized, at least at present, as a first pass over the data in order to expedite the process of identifying cognates or, in this case, loan words. Only needing to examine 716 lexical items and 232 concepts across all five lists instead of the total 7,358 lexical items and 1,301 concepts saves time.

The output of each hypothesis is a file containing items that fit the lexical distribution organized by concept. For the purpose of investigating the hypotheses, I went through each file and categorized each concept with a code to indicate its status as loan with origin of borrowing, cognate, or unknown. The overall results can be seen below:

Table 7: Overall Results

Result	Count	Percent
Not Loan	90	31.80%
Internal Loan	52	18.37%
Erroneous	20	7.07%
Loan: Arabic	15	5.30%
Loan: Unknown	12	4.24%
Loan: Russian	10	3.53%
Loan: Persian	9	3.18%
Unknown	7	2.47%
Loan: Turkic	6	2.12%



Table 7, continued

Duplicate	5	1.77%
Loan: Northeast Caucasian, undetected	4	1.41%
Loan: Iranian	2	0.71%

---

Overall, 90 concepts were determined to contain items that were most likely cognates, meaning each item was the result of direct inheritance and not borrowing. 52 concepts contain items that were determined to be the result of family-internal loans. 20 concepts contain items that were likely erroneously clustered together due to surface similarities that are not the result of either direct inheritance or borrowing. Fifteen concepts contained loanwords likely from Arabic; it makes sense that this is the largest source of external borrowing on this list, as Arabic was the one source of external borrowing that could not be filtered through a dictionary from the Intercontinental Dictionary Series. Twelve concepts contain items that are likely loanwords, but it is not clear the direction or source of the items. Ten concepts contain items that are likely the result of borrowing from Russian and nine from Persian. Seven concepts contain items that could not sufficiently be determined to be loanwords or the result of inheritance. Six concepts contain items likely to be borrowed from a Turkic languages (Kumyk or Azerbaijani). Five concepts are duplicates because these languages in question had the same term for two concepts (i.e., ‘daughter-in-law of a woman’ and ‘daughter-in-law’ of a man’). Two concepts contain items likely to be loaned from an Iranian language, but not necessarily Persian. Four concepts contain items that are in fact the result of family-internal (or Northeast Caucasian) loans but were not grouped that way because the specific lexical item borrowed was not included in the lexical list for the language it was borrowed from.

One immediate and possible interpretation of these results is that the LexStat program performed poorly. It is designed to detect true cognates (List, 2012a), meaning items that are descended from the same proto-word through direct genetic inheritance, and yet only 31.80% of the 283 concepts detected as cognate and examined here are true cognates. However, if examined from an alternate lens, the program may have performed fairly well. Only 7.07% of the 283 concepts investigated here were determined to be erroneously connected; therefore, 92.93% of items clustered together were actually related in some way, whether they be true cognates or loanwords. This emphasizes the importance of the computer-assisted approach. If the cognate detections for these concepts was assumed to be true, many loanwords would be marked as cognate; however, if one works off of the assumption that the items detected may be cognates or loanwords, the program can do a good job of bringing to the forefront related items for further investigation.

An example of a set that was correctly identified as cognates resulting from direct, genetic inheritance can be seen below:

Table 8: Set associated with 'navel'

Language	Alignment
Avar	c' i n u
Archi	c' a n -
Lak Shali	c' u n -

For this set, LexStat correctly identified these items as true cognates. Based on the cognate correspondences from Nichols (2003) and following her method of reconstructing solely consonants, the proto-Northeast Caucasian word for 'navel' could possibly be reconstructed as *\*c: 'Vn*.

An example of an erroneous clustering can be seen below:

Table 9: Set associated with ‘spring’

Language	Alignment
Avar	- - - i x
Archi	ī a n a q

For this set, the [ix] in Avar is aligned with the [aq] portion of the word [ī a n a q] in Archi. However, there is no reason to expect that these items are either cognates or loanwords based on the surface similarity of two phonemes in each item. As such, the items were erroneously clustered together.

An example of an undetected internal loan can be seen below:

Table 10: Example of an undetected internal loan

Language	Alignment
Avar	q <sup>w</sup> i l
Archi	q <sup>w</sup> i l

The Archi form here is clearly a loan from Avar. However, it was not clustered with Avar and did not appear under hypothesis two for one simple reason: the Avar form [q<sup>w</sup>il] was not included under the concept ‘bunch’ like Archi [q<sup>w</sup>il] was, or included in the Avar data at all. According to Chumakina (2009b), this form can mean ‘bunch’ as in the Intercontinental Dictionary Series, or ‘vine’, which is not a concept in the Intercontinental Dictionary Series and as such was left out (2023).

Additionally, while it makes sense that loans from Arabic could not be detected due to the reasons discussed above, there were some Persian, Russian, and Turkic loans that were not filtered out due to the similar fact that either the item that was borrowed was

grouped under a different concept or the concept of the external item was not included in the data at all. For instance, LexStat grouped the Archi word for ‘jaw’, [čarx], with the Avar word [x<sup>w</sup>enex]. However, Chumakina (2009a) notes that this word in Archi is a borrowing from the Persian word for ‘spool’. As ‘spool’ is not included in the concept list for the Intercontinental Dictionary Series (2023), this would never have appeared as a possibility in any cognate or loan detection program.

This brings up one of the key issues with many computational approaches to cognate and borrowing detection, which is how to organize the data and how to handle the notion of semantic drift and synonyms. While the Intercontinental Dictionary Series (2023) does allow for synonyms, broadening its reach and narrowing the chances of having the “wrong” synonym leading to an undetected cognate or borrowing, it obviously does not contain items entirely outside of the semantic reach of its concept list. Many linguists have sought to make the concept lists as small as possible in order to have the most conservative, most resistant-to-borrowing concepts for comparison and avoid any “noise” that may come from having a larger, more easily replaceable concept list (Heggarty, 2010, p. 317). However, Heggarty (2010) argues that there is no reason to throw away data from less stable concepts, as these concepts can be highly informative about the relationship between languages, especially in comparison to more stable concepts.

With regards to synonyms, there is some contention between whether synonyms should be encouraged or discouraged. Heggarty (2021, p. 389) argues that the concepts within a concept list should be as specific and narrowly defined as possible; at least for phylogenetic purposes, consistency within definitions is what should be most important. However, others have sought to account for and include semantic drift. For

instance, Kondrak (2009) created a cognate detection system that ignores concepts when detecting items such that two items with different glosses can be compared to one another. The items are then automatically evaluated for semantic similarity considering glosses, glosses and keywords, or WordNet relations, with more similar items deemed more likely to be cognate. In the LingPy system, on the other hand, items are only compared to one another within concepts (List & Forkel, 2021). As a result, items that have experienced semantic drift cannot be directly compared to one another within the cognate detection system; human knowledge would need to be applied later to determine that the items are cognates or loanwords that have undergone some semantic drift.

Another issue is with the goal of the computational approaches versus the output. LexStat is described as a cognate detection program, and while non-loans had the highest count in the filtered results overall, at 88 out of 232, that is not even the majority of the items. 54 of the items were determined to likely be internal loans, 20 were erroneous alignments, and 58 were loans from other sources. The total number of loans detected as cognate was therefore 112, as compared to the 88 non-loans. I believe these results reemphasize the critical idea that computational approaches for cognate and loan detection may be better referred to or treated as computer-assisted approaches. The program did a decent job of finding words that were similar for various reasons; only 20 out of 232, or 8.6%, of items were determined to be erroneous. However, each item needed to be investigated to determine if it was actually cognate or a loanword, and if it was a loanword, from where.

Essentially, the “cognate detection” program cannot truly identify cognates; it does not know what a cognate is. It is detecting items that are phonetically similar within a

certain threshold and contained within the same meaning slot. LexStat does attempt to determine sound correspondences (List, 2012a, p. 120), a key step to identifying cognates, but loanwords from one language to another often undergo phonological adaptation, which can have the appearance of regular correspondences. Additionally, while we may expect cognates to be somewhat phonetically similar to each other, so are loanwords. In fact, words that are too similar may be more likely to be loanwords than cognates if the languages in question are hypothesized to have diverged earlier in the history of the language family, and may be practically guaranteed to be cognates if the languages in question are from different families (an exception would be for random chance similarities). Taking the cognate judgements at face value would have meant interpreting many loanwords as the result of genetic inheritance. Overall, the results clearly show that these systems are not perfect; interpretation is critical. Each individual hypothesis is discussed below, and the possible loanwords from Avar and Lak are treated individually in section 4.2.1.

For the first hypothesis, loanwords from Avar into Archi and Lak Shali, the results in table form are blow:

Table 11: Results of Hypothesis 1

Result	Count
Internal Loan	2
Loan: Arabic	2
Not Loan	1
Loan: Russian	1
Loan: Turkic	1
Loan: Unknown	1

For hypothesis one, out of the eight concepts detected as similar, three were determined to be internal loans of the type being sought, meaning from Avar into Archi and/or Lak Shali. The rest were loans from external sources, including Arabic, Russian, an unknown Turkic language, and an unknown source.

Hypothesis two sought loanwords from Avar into Archi alone, and had far more concepts to investigate than the hypothesis one: 140 compared to eight. The results of this hypothesis are below:

Table 12: Results of Hypothesis 2

Result	Count
Not Loan	55
Internal Loan	39
Erroneous	13
Loan:Arabic	7
Loan: Unknown	6
Loan: Persian	5
Loan: Russian	4
Duplicate	5
Unknown	3
Loan: Turkic	2
Loan: Iranian	1

As can be seen from this table, out of the 140 concepts, 55 of them were determined to be cognates or likely cognates, meaning they are perceived as more likely to be the result of direct inheritance than borrowing. 39 of them were determined to likely be internal loans, meaning loans from Avar to Archi, and 13 of them were determined to be the result of

erroneous similarity detection. By erroneous similarity detection, I mean LexStat clustered them together, but they appear to be neither cognates nor loans, but rather items with some small level of surface similarity. Seven of the items were determined to be loans from Arabic, six were determined to likely be loans but from unknown sources, five were determined to be loans from Persian, four were determined to be loans from Russian, three were labeled as unknown, meaning either loan or cognate, two were determined to be loans from a Turkic language, and one was determined to be a loan from Iranian.

Hypothesis three was intended to determine borrowings from Avar into Lak Shali, and contained items across 13 concepts. The results for hypothesis three are below:

Table 13: Results of Hypothesis 3

Result	Count
Erroneous	6
Loan:Arabic	3
Unknown	2
Internal Loan	1
Loan: Unknown	1

Only one of the concepts from this hypothesis was determined to be an internal loan of the type being sought. Most of the items detected as similar were erroneously connected. Three concepts are the result of loans from Arabic, one is likely to be a loan from an unknown external source, and two are unknown.

Hypothesis four was intended to determine loans from Lak into Archi and contained 91 concepts. The results of the output from hypothesis four can be seen below:



Table 14: Results of Hypothesis 4

Result	Count
Not Loan	28
Internal Loan	10
Loan: Russian	4
Loan: Persian	4
Loan:Arabic	3
Loan: Unknown	3
Loan: NE, Undetected	2
Unknown	2
Loan: Turkic	2
Loan: Iranian	1

For this hypothesis, 28 items were determined to be more likely to be the result of direct inheritance, meaning cognates, as opposed to loans. Ten were determined to be internal loans of the type being sought. Four each were determined to likely be loans from Russian and Persian, and three each were determined to likely be loans from Arabic and an unknown source. Two were determined to be internal loans; however, this determination was not made based on the similarity detection of the items but rather through later investigation. As such, these items are listed as loans from North East Caucasian that were undetected. In both cases, these loans were undetected because the item within the language family that they were borrowed from was not part of the dataset that was fed into LexStat. Lastly, two loans were determined to likely be from a Turkic language, and one was determined to likely be from an Iranian language.

The final hypothesis under review is hypothesis five, possible loans from Archi in Lak Shali, which contains twelve concepts. The results for this hypothesis can be seen below:

Table 15: Results of Hypothesis 5

Result	Count
Not Loan	6
Loan: NE, Undetected	2
Loan: Russian	1
Loan: Unknown	1
Erroneous	1
Loan: Turkic	1

This was the only hypothesis to not contain a single internal loan detected by the system. Six of the concepts were determined to likely contain items that are the result of direct inheritance while two were likely loans from Northeast Caucasian languages that went undetected, one was likely a loan from Russian, one was likely a loan from an unknown source, one was likely a loan from a Turkic language, and one was likely an erroneous grouping.

Overall, the sets examined do reveal possible lexical borrowing between Avar, Archi, and Lak. However, there were still many loanwords of uncertain origin or direction, loanwords from Turkic (most likely Kumyk or Azerbaijani), Persian, Arabic, Iranian, and Russian that were detected even with their data being included to filter them out, and words with only surface resemblance erroneously detected by the system.

## 4.2.1 Transcription Style

For the transcription, the following conventions will be used:

Table 16: Transcription Conventions

Symbol	Sound
š	Voiceless palato-alveolar fricative, [ʃ]
ž	Voices palate-alveolar fricative, [ʒ]
c	Voiceless alveolar affricate, [ts]
č	Voiceless palato-alveolar affricate, [tʃ]
č̣	Voiceless alveolar lateral affricate, [tʃ̣]
y	Voiced palatal approximant, [j]
’	Ejective marker
–	Length marker (e.g., [c̄] = [c:] = [cc])

The results section below will discuss in detail all sets of lexical items possibly indicating borrowing amongst these languages, beginning with loans from Avar into both Archi and Lak Shali, from Avar into Lak Shali, and from Avar into Archi, and finally loans from Lak into Archi. Unless otherwise noted, all lexical data is from the Intercontinental Dictionary Series (2023).

## 4.2.2 Possible Borrowings from Avar

The first three hypotheses represent loanwords from Avar into Archi and Lak Shali. The lexical items with a distribution such that they appear in Archi, Avar, and Lak Shali but not standard Lak are likely to be borrowings from Avar into Archi and Lak Shali; additionally, there are words that appear only in Avar and Archi, and only Avar and Lak Shali. Each item will be discussed in some detail below. Note that each item is given in

its alignment form with gap characters (-) inserted as needed in order to align the sounds being compared; any dash within the tables is therefore not intended to represent a morpheme boundary.

Table 17: Set associated with ‘booty, spoils’

Language	Alignment
Avar	d a w l a
Archi	d u w l i
Lak Shali	d a w l a

For this set, the fact that the Avar form and Lak Shali form are identical, [dawla], is a strong indication of borrowing in some direction. Avar and Lak Shali are from different branches of the Northeast Caucasian family tree and therefore one would expect more change between the forms over time if they represented direct inheritance from an original proto-form. The Archi form is also similar, [duwli], with [u] instead of the first [a] and [i] instead of the second [a]. However, it is not clear whether this word is a loanword from Avar into Archi or the result of a loanword from a non-Northeast Caucasian language into Archi, Lak, and Avar. Chumakina (2009b, p. 440) states that vowel-raising is a common phenomenon for words borrowed from Avar into Archi, meaning it may be expected to get [duwli] from [dawla]. the initial [a] would raise to [u], with labialization perhaps the result of the following [w], and the last [a] would raise to [i]. Other Avar-Andic-Tsezic languages as well as Lak Balkhar and some dialects of Dargwa have the form [dawla] as well, and some other Lezgian languages, of which Archi is a member, such as Lezgian Mirakh and Southern Tabasaran, have [dewlet] as their form. It is also possible this item is a loanword into Avar and by extension Archi from Arabic.

Table 18: Set associated with ‘heart’

Language	Alignment
Avar	r a k’
Archi	- i k’ <sup>w</sup>
Lak Shali	d a k’

For this set, LexStat correctly identified Avar [rak’] and Archi [ik’<sup>w</sup>] as true cognates; these are each descended from the proto-Northeast Caucasian [rVk’u] or [Vrk’u] proposed by Nichols (2003, p. 258). However, [dak’] is not what would be expected for Lak Shali if that item was the result of direct inheritance: that would be [qqʁuk’] or [qqu<sup>ʁ</sup>k’] (Nichols, 2003, p. 258). As such, the Lak Shali form is possibly the result of borrowing or contact from Avar. The only other forms beginning with [d] for ‘heart’ are from the Nakh languages, Ingush [dog], Chechen [dog], and Tsova-Tush [dok’] (*The Intercontinental Dictionary Series*, 2023). More correspondences showing [r] in Avar and [d] in Lak Shali that are clearly loanwords or cognate would be useful in determining the status of this item more definitively.

Table 19: Set associated with ‘stinking, bad-smelling’

Language	Alignment
Avar	m a ħ c e l
Lak Shali	m a ħ - - -

For this set, only this initial portion of the word is the same: [maħ]. This portion in Avar may be a more neutral word for smell, as evidenced by the intransitive verb [maħ buk’ine] ‘smell’ (verb, intransitive) where [buk’ine] also means ‘to have’ (Khalilov,

2023c). The standard Lak item is [cuḡannu], which is of completely different origin; as such, the Lak Shali item is possibly the result of contact with Avar.

Table 20: Set associated with ‘Friday’

Language	Alignment
Avar	r u z m a n
Archi	r u z m a n

For this set, the items are identical, strongly indicating a borrowing or contact-related relationship. Other Avar-Andic-Tsezic languages also show [ruzman], such as Bagvalal and Andi, while Chamalal shows [ruzmä] (*The Intercontinental Dictionary Series*, 2023). On the other hand, many Lezgian reflexes demonstrate a different pattern likely to be cognate with the Avar forms, such as the Rutul [žuma], Tabasaran [žʷumi], and Lezgian [žümya] (*The Intercontinental Dictionary Series*, 2023). As Archi is a member of the Lezgian branch of the Northeast Caucasian family, one would expect it to have a similar reflex beginning with a voiced alveolar affricate or fricative. Instead, since its form is identical to the Avar form and other Avar-Andic-Tsezic reflexes, this similarity is likely due to contact effects.

Table 21: Set associated with ‘arrow’

Language	Alignment
Avar	č' o r
Archi	č' o r

For this set, the Avar and Archi items for ‘arrow’ are identical, indicating some contact effects. If we examine the word for ‘arrow’ in other Lezgian languages to which Archi is more closely related, we see [x̣el] in Aghul and Lezgian as well as [ux] in some dialects of Rutul and Southern Tabasaran and [ox] in Khinalug and Budukh (*The*

*Intercontinental Dictionary Series*, 2023). Examining the consonant correspondences from Nichols (2003) does not show a set in which [čʼ] in Avar would correlate with [ǰ] in Aghul and Lezgi or a null character in Rutul, Tabasaran, and Khinalug (Budukh is not present in the correspondences), meaning perhaps these words are not etymologically related back to proto-Northeast Caucasian. They may be from different roots originally that have converged on the same meaning due to semantic drift, or the consonant correspondences may be disguised due to heavy borrowing. Alternatively, as [r] and [l] are both liquids and [ǰ] and [čʼ] are both voiceless consonants, it could also be possible for them to be descended from a form beginning with a velar voiceless ejective consonant that was perhaps palatalized in some varieties and lost its ejective form in others. Overall, for this term, the Archi item [čʼor] does seem to possibly be the result of contact with Avar.

Table 22: Set associated with ‘bow’

Language	Alignment
Avar	čʼ o r b u tʼ
Archi	čʼ o r b u t

For this set, the forms are identical with the exception of the final consonant: Avar has the ejective [tʼ] while Archi has the plain [t]. However, the Archi form ‘bow’ is given as [čʼorbutʼ] with the final ejective [tʼ] in the Dictionary of Archi (Chumakina et al., 2007). These forms also appear obviously related to the forms for ‘arrow’ given above with the additional ending [but].

Table 23: Set associated with ‘bark’

Language	Alignment
Avar	q a l
Archi	q a l

For this set, the Avar and Archi items are once more exactly identical. This indicates borrowing as there are no correspondence sets in Nichols (2003) that would predict a [q] in both Avar and Archi if both items were the result of direct inheritance.

Table 24: Set associated with ‘blister’

Language	Alignment
Avar	p u l
Archi	p i l

For this set, the Avar and Archi items contain the same consonants but differ in vowel quality. There do not seem to be many related words in the set from the Intercontinental Dictionary Series; the Avar-Andic-Tsezic languages Southern Akhvakh and Godoberi have [pale] and [puli] respectively and Karata has [pɪʎʎ’a], but other languages in the same branch appear to have etymologically unrelated words (*The Intercontinental Dictionary Series*, 2023). For instance, Chamalal has [čičil], Andi has [č’akara], and Bagvalal has [čix̄<sup>w</sup>]. In the Lezgian branch of which Archi is a member, there again appears to not be a word directly related to [pul]. Udi has [to<sup>s</sup>f], Tsakhur has [gabar], Lezgian has [kurkur], and Northern Tabasaran has [k’aš] (*The Intercontinental Dictionary Series*, 2023). The Mirakh dialect of Lezgian (Lezgian branch) does have [pelex], which is the only form in the Lezgian branch that seems similar to the Avar and



Archi reflexes. The consonant correspondences set up by Nichols (2003, p. 247) have [p] in Archi correlating to [p] in Avar, which would match the reflexes given; however, the fact that this form seems more present in the Avar-Andic-Tsezic branches indicates that the form in Archi is perhaps due to contact, either as borrowing from Avar or reinforced by the similar form in Avar, if the evidence from [pelex] in Mirakh Lezgian is interpreted to mean that a [pVI] form also existed in Lezgetic and therefore in Archi.

Table 25: Set associated with ‘blood’

Language	Alignment
Avar	b i
Archi	b i

For this set, both the Avar and Archi items are identical. This set is interesting, as there are not a lot of items that appear to be descended from the same root in either the Lezgetic or Avar-Andic-Tsezic branches. In a Swadesh list of lects of the Lezgetic branch, Kassian (2011) states that there are two proto-Lezgetic roots that are used to mean ‘blood’ in the Lezgetic branch: *\*p:iy* and *\*ʔäʔ*. The reflexes of *\*p:iy* exist in the Lezgetic branch as ‘blood’ only in Archi [bi] and Udi [p:i]. He proposes that *\*ʔäʔ* originally meant ‘blood’ with its reflexes still meaning blood in many Lezgetic languages, and that *\*p:iy* meant ‘blood vessel’ or ‘vein’ before shifting to the meaning ‘blood’ in both Udi and Archi independently. The Khinalug word for ‘blood’ is [p’i]. Khinalug is sometimes placed within the Lezgetic branch, meaning it would then also need to have independently shifted the meaning of *\*p:iy* to blood, or it is sometimes placed as a branch-level isolate within the family.

If the proto-Lezgian root for ‘blood’ is taken to be *\*ʔäʔ* as argued by Kassian (2011), that makes it interesting that Avar also has [bi] for ‘blood’. Kassian (2013) proposes *\*ħǝy* as the proto-Tsezic root for ‘blood’, and the Tsezic languages are frequently grouped with the Avar-Andic languages (Dobrushina et al., 2020b, p. 30). Within the DagSwadesh database, a collection of Swadesh lists collected on the village level from languages considered to be members of the Avar-Andic branch, every lect contains words seemingly possibly related to the Tsezic root except Avar’s [bi] (Filatov & Daniel, n.d.). For instance, Godoberi has [hiri], Tukita has [hini], and Tindi has [heri]. As such, Avar’s form appears to be unique among the Avar-Andic-Tsezic languages.

Lastly, if we look at the Dargic and Lak languages, we see the potential for the two different roots once more. In Belyaev (2014), the forms for ‘blood’ from three Dargic lects are given as [biʔi] (Shiri), [beʔ] (Amuzgi) and [bay] (Ashti) and the proto-Dargwa form *\*beħ(i)* is suggested. The Lak forms, on the other hand, are given as [oʃ] for standard Lak, Lak Shali, and Lak Arakul, and [oʃtu] in Lak Balkhur (*The Intercontinental Dictionary Series*, 2023). The pharyngealization indicates that the proto-Lak root is likely related to the proto-Tsezic *\*ħǝj* and proto-Lezgian *\*ʔäʔ*, whereas the proto-Dargwa root is likely related to the proto-Lezgian *\*p:iy* and the Avar form [bi].

Overall, if the root for Avar [bi] and Archi [bi] is taken to originally mean ‘blood vessel’ (Kassian, 2011), this would mean that the meaning shifted to ‘blood’ in Avar (Avar-Andic-Tsezic), Archi (Lezgian), Udi (Lezgian), Dargic, and Khinalug (Lezgian or isolate), meaning this shift would have happened possibly independently in multiple branches: Avar-Andic-Tsezic, Lezgian, Dargic, and Khinalug (if treated as outside of the

Lezgi branch). It is also possible that the Archi form [bi] is the result of contact from Avar, either as direct borrowing or as influence, especially considering the Udi form is [p:i]. While the consonant correspondences compiled by Nichols (2003) do not contain correspondences for the sound [p:] in Udi, she does align Archi [b] with Udi [b] (2003, p. 249). More correspondence sets would be needed to determine if Udi [p:] shows regular correspondence with Archi [b] in order to determine the likelihood that the Archi form is the result of direct inheritance from Proto-Northeast Caucasian or contact influence from Avar.

Table 26: Set associated with ‘boundary’

Language	Alignment
Avar	ʃ u r q i
Archi	ʃ u r q i

For this set, the Archi and Avar forms are identical. Lezgi forms include Lezgian [časpar], Budukh [serhet], Southern Tabasaran [saʳrhaʳt], and Rutul [žabsar], while Avar-Andic-Tsezic forms include Andi Muni [orqi], Botlikh [ʃurqi], Hunzib [ʃorqi], and Chamalal [orqi]. As the Archi form is identical to the Avar form and seemingly unrelated to any of the other Lezgi forms, it is likely that this form is the result of contact with Avar.

Table 27: Set associated with ‘cattle’

Language	Alignment
Avar	b o č' i
Archi	b u č' i

For this set, the items for ‘cattle’ in Avar and Archi are identical with the exception of the first vowel, which is [o] in Avar and [u] in Archi. This would fit the pattern of borrowing, as Chumakina (2009b, p. 440) states that loans from Avar to Archi can undergo vowel raising, as in [o] to [u]. Additionally, while the consonant correspondences established by Nichols (2003) do show a correspondence between [b] and [b] in Avar and Archi, a [c̄ʰ] in Avar would be expected to correspond with [cʰ i] in Archi. Instead, in this set the same consonant is used for both. As such, the Archi word [bučʰi] fits the profile of what would be expected given borrowing from Avar.

Table 28: Set associated with ‘cock, rooster’

Language	Alignment
Avar	ħ e l e k o
Archi	ħ e l e k u

For this set, the items for ‘rooster’ are again identical with the exception of a vowel, in this case the [o] in Avar [ħeleko] and the [u] in Archi [ħeleku]. This would again fit the pattern of potential vowel raising in Archi borrowings from Avar (Chumakina, 2009b, p. 440), with the [o] raising to [u]. Chumakina (2009b, p. 444) includes the Archi word [ħeleku] in her list of loanwords from Avar into Archi as well.

Table 29: Set associated with ‘daughter-in-law’

Language	Alignment
Avar	n u s - - -
Archi	n u s d u r

For this set, the beginning of each item, [nus] is identical for both Archi and Avar. Chumakina (2009b, p. 444) states that this is a loanword from Avar, and that the

[-du-] is an adjectivizing suffix and the final [-r] is a gender marker. According to Chumakina (2016, pg. 3597), [-du-] is an allomorph of the suffix [t̄u]. Therefore, overall it appears likely that the [nus] form from Avar has been borrowed into Archi and then given Archi morphology.

Table 30: Set associated with ‘ditch’

Language	Alignment
Avar	r a q
Archi	r a q

For this set, the Archi and Avar items are identical. While Nichols (2003, p. 254) does state that [r] can correspond for Archi and Avar, the other Lezgian languages have a different form for this item that appears to be related, implying metathesis of the vowel and liquid in one of the ancestor branches: [arx] (Southern Tabasaran, Udi, Aghul Koshan, Tsakhur, and Lezgian Mirakh), and [erx] (Budukh). On the other hand, [raq] and other similar forms are more prevalent in the Avar-Andic-Tsezic languages: [raq] (Avar), [raq̄e] (Northern Akhvakh), [roqi] (Andi Muni), [ruhi] (Bezhta Khasharkota), [reqin] (Botlikh, Godoberi, Karata) (*The Intercontinental Dictionary Series*, 2023). Since the Archi form is identical to the Avar form and dissimilar from the form that would be expected given its placement in the Lezgian branch, it is likely that this form is borrowed from Avar. Chumakina (2009b, p. 444) also labels this item in Archi as a loanword from Avar.

Table 31: Set associated with ‘east’

Language	Alignment
Avar	b a - q' _ b a - - - k̄ u l _ r a q
Archi	b a r q _ b o r λ i n n u t _ r a q

This set is indicated as a borrowing into Archi from Avar in Chumakina (2009b, p. 444). The form [raq] at the end of the phrase has already been stated above to be a borrowing from Avar.

Table 32: Set associated with ‘family’

Language	Alignment
Avar	q i z a n
Archi	q i z a n

For this set, the Archi and Avar forms are identical. Many of the Lezgetic lects show the form [xizan] with the [x] phoneme instead of [q] (ie., Tsakhur, Tabasaran, and Aghul) while the [qizan] form is more common in the Avar-Andic-Tsezic lects (*The Intercontinental Dictionary Series*, 2023). While this does not guarantee borrowing, as it would be possible for the [q] to be the original sound that changed to [x] in most of the Lezgetic lects but not Archi, the consonant correspondence sets in Nichols (2003, p. 249) indicate that a [q] in Archi would be expected to correspond with a [qʼ] in Avar. There is not a suggestion for what sound in Archi the [q] in Avar would correspond with. As such, this item in Archi may be borrowed from Avar, or the [q] may have been preserved in Archi due to contact with Avar, but more cognate sets would be needed to confirm the expected correspondences.

Table 33: Set associated with ‘father-in-law’

Language	Alignment
Avar	w a λ λʼ a d
Archi	w a λ λʼ a d

For this set, the Archi and Avar forms are identical. While the consonant correspondences in Nichols (2003) do contain correspondences for Avar and Archi [ʎʎ'] or [w], Chumakina (2009b, p. 444) does identify this item as a loanword from Avar.

Table 34: Set associated with ‘fermented drink’

Language	Alignment
Avar	č' a ʎ a
Archi	č' a ʎ a

For this set, the Avar and Archi forms are identical. Other Avar-Andic-Tsezic languages have similar forms with the internal pharyngeal, such as [č'eʎe] (Bagvalal) and [č'aʎa] (Godoberi, Hunzib, Bezhta, Karata, Khwarshi), while other Lezgian languages have an internal [x], such as [čexir] (Mirakh Lezgian), [čaxir] (Aghul), and [čaxir] (Rutul). While the established consonant correspondences in Nichols (2003) do indicate that [č'] would correspond in Avar and Archi, the same correspondences also indicate that [x] in Lezgian would be expected to pattern with [x] in Archi (and Avar). However, Nichols (2003) does not contain correspondences for pharyngeals in general, so it is possible that an [ʎ] in Avar-Andic-Tsezic languages could correspond with an [x] in Archi. Regardless, it seems likely that the differences in the Archi form from other Lezgian forms may be the result of contact with Avar.

Table 35: Set associated with ‘fisherman’

Language	Alignment
Avar	č̃ u ʎ i q a n
Archi	č̃ u ʎ i q a n

For this set, the Avar and Archi forms are once again identical. The consonant correspondences from Nichols (2003, p. 239) suggest that a [q] in Archi would be expected to correspond with a [qʼ] in Avar if both were the result of direct inheritance instead of the [q] seen (there is not a correspondence set demonstrating what would be expected given a [q] in Avar). Based on this mismatch of correspondences as well as the fact that this fairly long word is identical in two languages that diverged likely thousands of years ago, it seems most likely that the word in Archi is the result of contact with Avar.

Table 36: Set associated with ‘forehead’

Language	Alignment
Avar	n o d o
Archi	n o d o

For this set, the Avar and Archi forms are identical, indicating borrowing or contact effects. Additionally, Chumakina (2009b, p. 444) marks this item as a loan from Avar.

Table 37: Set associated with ‘hammer’

Language	Alignment
Avar	k <sup>w</sup> a r t’ a
Archi	k’ u r t’ a

For this set, the Avar and Archi items are similar but not identical. The first vowel is [a] in Avar and [u] in Archi; this could be explained by the fact that many loanwords from Avar into Archi demonstrate vowel-raising (Chumakina, 2009b, p. 440). The difference in the initial consonant is more interesting. Avar has [k<sup>w</sup>] which is a labialized voiceless velar stop, while Archi has [k’], an ejective voiceless velar stop. Some other dialects of Avar show [kurt’a], with an initial plain voiceless velar stop (*The*



*Intercontinental Dictionary Series*, 2023). In this case, it would be useful to know the exact form utilized in the village of Chilab and its initial consonant. Chumakina (2009a) labels this item in Archi as ‘potentially borrowed’.

Table 38: Set associated with ‘idol’

Language	Alignment
Avar	q a n č
Archi	q a n č

For this set, the Avar and Archi forms are identical. While the consonant correspondence sets in Nichols (2003) do not demonstrate correspondences for [q], they do indicate that a [č] in Avar would be expected to correspond with a [š] in Archi and a [č] in Archi would be expected to correspond with a [c’] in Avar. As such, since the consonants in Archi and Avar do not correspond in the way that we would expect given direct inheritance for both, it is likely that the form in Archi is the result of borrowing or contact with Avar.

Table 39: Set associated with ‘island’

Language	Alignment
Avar	č’ i n k’ i l l i
Archi	č’ i n k’ i l l i

For this set, the Avar and Archi forms are identical. Given the fact that Avar and Archi are in separate branches of the Northeast Caucasian language family and would have diverged long ago, it is likely that the form in Archi is the result of contact with Avar. Additionally, the consonant correspondences in Nichols (2003, p. 248) suggest that while a [č’] in Avar can correspond with either a [č’] or [c’] in Archi, the [l] in Avar is

expected to correspond with [H] in Archi. Overall, due to this mismatch of correspondences as well as the unexpected similarity in forms, it is likely the form in Archi is the result of contact with the form in Avar.

Table 40: Set associated with ‘keep, retain’

Language	Alignment
Avar	c’ u n i z e
Archi	c’ u n a s -

For this set, the Avar and Archi items are similar; the Avar form ends in [ze] while the Archi form ends in [s]. As both word-final devoicing and intervocalic voicing are common phonological processes generally, this difference in ending is not unexpected. Additionally, according to the consonant correspondences in Nichols (2003), a [c’] in Avar would be most likely to correspond with either a [c] or [č] in Archi while a [c’] in Archi would be most likely to correspond with a [č’] in Avar. Chumakina (2009a) states that the item [c’unas] in Archi is ‘potentially borrowed’, and based on the fact that the consonant correspondences are not what would be expected given direct inheritance, it seems likely that this item is the result of contact with Avar.

Table 41: Set associated with ‘leather’

Language	Alignment
Avar	q a l
Archi	q a l

For this set, the items in Avar and Archi are identical. According to the consonant correspondences in Nichols (2003, p. 249), a [q] in Archi would be expected to align with

a [q'] in Avar. Due to this mismatch in expected correspondences, it is possible that the item in Archi is the result of contact with Avar.

Table 42: Set associated with 'lion'

Language	Alignment
Avar	γ a l b a c'
Archi	γ a l b a c'

For this set, the Archi and Avar items are identical. While there are not correspondences established in Nichols (2003) between Archi and Avar for [γ], the length of the item and its identicalness across both languages is highly indicative of borrowing. Given that Archi and Avar are members of separate branches of the Northeast Caucasian family that diverged long ago, it would be highly unusual for these two items to have remained the same in that time.

Table 43: Set associated with 'lip'

Language	Alignment
Avar	k'w e t'
Archi	k'w e t'

For this item, again the Archi and Avar items are identical. According to the Global Lexicostatistical Database (Starostin, 2011), the expected Archi form would be [k'went'] if the item in Archi was the result of direct inheritance, with a [n] preceding the [t']. The other Lezgian lects with an item that seems related to this form are Aghul with [k'ent'w] and Southern Tabasaran with [k'want'] (*The Intercontinental Dictionary Series*, 2023). As the [n] is missing and the form is identical for Avar's form, it is possible that the item is

borrowed from Avar, or that long-term contact with Avar has caused the Archi form to appear more like that Avar form than the other Lezgian forms.

Table 44: Set associated with ‘live, living, life’

Language	Alignment
Avar	č’ a g o y a b -
Archi	č’ a g u t̄ u

For this item, the Avar and Archi items are similar. The [-yab] in the Avar form is a combination of the suffix [-ya-] used to derive adjectives from nouns and the gender marker [-b] (Khalilov & Khalilova, 2016, p. 3701), and the [t̄u] in the Archi form is an adjectival derivational suffix (Chumakina, 2009b, p. 441). As such, with the additional morphology removed, the [č’agu] segment of the Archi item alone is to be compared with [č’ago] in Avar. As discussed in Chumakina (2009b, p. 440) and mentioned previously, Archi sometimes raises vowels of loanwords borrowed from Avar, which would account for the [u] compared to [o]. While [č’] can correspond between Archi and Avar (Nichols, 2003, p. 248), the only other form that appears potentially related among the Lezgian languages is [č’iwir] (Southern Tabasaran) (*The Intercontinental Dictionary Series*, 2023). Therefore, it seems likely that the base form [č’agu] in Archi was borrowed from Avar [č’ago]. The fact that just the base form was borrowed without the Avar suffix indicates that Archi is able to integrate this Avar loanword into its morphological system as well (Chumakina, 2009b, p. 441).

Table 45: Set associated with ‘magic, witchcraft, sorcery’

Language	Alignment
Avar	m a k r u t̄ i
Archi	m a k r - - u

For this set, the beginnings of the Avar and Archi items are identical. Interestingly, the alignment algorithm used as part of LexStat chose to align the [u] in the Archi item with the final [i] in Avar instead of the [u] in [ru]. In Avar [-hi] is a suffix that “derives abstract and concrete nouns from nouns in the absolutive and oblique cases” (Khalilov & Khalilova, 2016, p. 3699). It is therefore more likely that the final [u] in the Archi item should be aligned with the [u] in [ru] in the Avar item than with the final [i]. Overall, given that the initial portion of the item [makru] is identical to Archi, which is unexpected given the time-depth separating Avar and Archi, it seems likely that the Archi item is the result of contact with Avar.

Table 46: Set associated with ‘mosquito’

Language	Alignment
Avar	$\bar{k}'$ a r a
Archi	k' a r a

For this set, the Avar and Archi items are identical except for the length of the initial consonant, with the Avar version being a long ejective [ $\bar{k}'$ ] and the Archi version being a short ejective [k']. However, this would be expected, as Archi does not have a long ejective [ $\bar{k}'$ ] as a part of its consonant inventory (Chumakina, 2020, p. 283). Therefore, replacing the [ $\bar{k}'$ ] in the Avar item with a [k'] in the Archi equivalent would be a logical choice, as it is the closest consonant to the original. Additionally, while there are a variety of items that seem etymologically related to the Avar form in the Avar-Andic-Tsezic lects, such as [ $\bar{k}'$ ara] (Andi, Botlikh, Chamalal, Karata), [š'ara] (Bagvalal), [ $\bar{k}'$ ara] (Tindi, Godoberi), and [k'ara] (Tsez), there are no such forms in the Lezgian branch other than the Archi

form. Therefore, it is likely that the Archi form [k'ara] is the result of borrowing from Avar.

Table 47: Set associated with ‘mother-in-law’

Language	Alignment
Avar	ya λλ' a d
Archi	ya λλ' a d

For this set, the Avar and Archi items are identical. Based on the consonant correspondences in Nichols (2003, p. 251), a [λλ'] in Avar would be expected to correspond with [λ'] in Archi, unlike the [λλ'] that is seen. Additionally, given the length of the item, it would be expected for the Avar and Archi items to have diverged if both were the result of direct inheritance. As such, it is likely that the Archi word is the result of borrowing from Avar.

Table 48: Set associated with ‘oar’

Language	Alignment
Avar	r a x̄ a n
Archi	r a x̄ a n

For this set, the Avar and Archi items are once again identical, indicating the possibility of borrowing. Additionally, while there are some similar forms in other Avar-Andic-Tsezic languages, such as [reχin] (Andi, Akhvakh) and [raχan] (Tindi, Sagada Tsez, Bagvalal) (*The Intercontinental Dictionary Series*, 2023), there are not any items in the Lezgian languages that would appear to be descended from the same root. This indicates that the items for ‘oar’ may have been borrowed into the Avar-Andic-Tsezic branch from another family or coined earlier in the history of the branch and then borrowed into Archi.

Table 49: Set associated with ‘plaintiff’

Language	Alignment
Avar	ʃ a r z a č i
Archi	ʃ a r z a č i

For this set, the Avar and Archi items are once again identical, indicating borrowing is likely a part of the history of these items in these languages. The initial pharyngeal consonant is interesting because it has been lost in many Lezgian languages, such as in [arzači] (Quba Lezgian, Mirakh Lezgian, Budukh, Udi) but also retained in others, such as in [ʃɤzarzakar] (Gelmets Tsakhur) and [ʃarzači] (Kryz) (*The Intercontinental Dictionary Series*, 2023). Different dialects of Dargwa either have the pharyngeal or don’t, and while many languages in the Avar-Andic-Tsezic branch contain the pharyngeal, some do not, such as [arzači] (Bezhta) (*The Intercontinental Dictionary Series*, 2023). Overall, it seems likely that the Archi item is the result of borrowing, as it is identical to the Avar word even though they are from separate branches. It is also a possibility that the item has been borrowed into the family from an external source, or that the Archi form has been maintained made to appear more similar to the Avar form through contact influence.

Table 50: Set associated with ‘queen’

Language	Alignment
Avar	b i k a
Archi	b i k a

For this set, the Avar and Archi items are identical. The Archi item is likely a loanword from Avar because the Lezgian branch shows [p] as the initial consonant for the majority of its forms and either internal [cc], [č] or [čč] instead of [k].

Table 51: Set associated with ‘ring (for finger)’

Language	Alignment
Avar	b a r ɣ i č -
Archi	b a - ɣ i ž a

For this set, the Avar and Archi items are similar. The final [č] consonant in Avar is matched with [ž] in Archi; this is not unexpected, as Chumakina (2009b, p. 440) states that loans from Avar to Archi sometimes undergo phonological adaptation in which Avar affricates, such as [č] in this case, become fricatives, such as [ž] in this case, in Archi. The [r] before the [ɣ] has also disappeared in Archi and an [a] has been added to the end. Critically, there do not appear to be any forms related to this one in the Lezgian languages but several in the Avar-Andic-Tsezic languages. Given the similarities between these two items in Avar and Archi as well as the explanation for the consonant change being the result of phonological adaptation, it is possible that this item in Archi is a loanword from Avar.

Table 52: Set associated with ‘sea’

Language	Alignment
Avar	r a ɫ a d
Archi	- - ɫ a t

For this set, there are some similarities in the Avar and Archi forms. The Archi form does not have the [ra] present at the beginning of the Avar form, and the final [d] in Avar is a [t] in Archi. The [d] to [t] change could be explained through word final devoicing, or a shift from [d] to [t:] to [t] (Starostin, 2011). It is not clear what would cause the loss of the [ra] segment from Avar to Archi, but given that the Lezgian languages have



forms that are seemingly related but different from the Avar form, such as [hül] (Lezgian), [huʕl] (Southern and Northern Tabasaran), and [hul] (Aghul), it seems likely that the Archi form is a loan from Avar.

Table 53: Set associated with ‘sorcerer, witch’

Language	Alignment
Avar	q a r t ay
Archi	q a r t ay

For this set, the Avar and Archi items are identical. Given that Avar and Archi diverged long ago, it would be unlikely for these items to still be identical. Additionally, according to Nichols (2003, p. 249), a [q] in Archi would be expected to correspond with a [qʰ] in Avar, which is not the case here. Since the items are unexpectedly similar and the correspondences are not what would be expected given genetic inheritance, it is likely that the item in Archi is the result of a loan from Avar.

Table 54: Set associated with ‘suspect’

Language	Alignment
Avar	š̄ a k ł i _ k̄ e z e
Archi	š̄ a k - - _ k e s -

For this form, there are similarities between the Avar and Archi forms. The underscore in the alignment represents a space, showing that each item is actually comprised of two individual words. The [š̄ak] element is likely borrowed from Avar as [š̄] in Archi would be expected to correspond with [x] in Avar (Nichols, 2003, p. 253). The [li] in the Avar item is a derivational suffix that derives “abstract nouns and concrete nouns from nouns in the absolutive and oblique cases” (Khalilov & Khalilova, 2016, p. 3699),

and as such it is likely that Archi borrowed the [šak] form from Archi without this suffix. The second portion of the item, [kes] in Archi, is a verb meaning ‘to become’ (Chumakina et al., 2007). Overall, it appears that the [šak] has likely been borrowed from Avar.

Table 55: Set associated with ‘taste’

Language	Alignment
Avar	t’ a ʕ a m
Archi	t’ a ɣ a m

For this set, the Archi and Avar forms are highly similar, more so than would be expected given the distance in time that these two languages have diverged. The Avar form has the pharyngeal consonant [ʕ] where the Archi form has the velar (or possibly uvular<sup>5</sup>) consonant [ɣ]. Chumakina et al. (2007) actually give the the Archi form for ‘taste’ as [t’aʕam], which is identical to the Avar form. Overall, given the unexpected similarities in the forms, it is likely that the Archi form is the result of borrowing from Avar.

Table 56: Set associated with ‘thief’

Language	Alignment
Avar	c’ o h o r
Archi	c’ o h o r

<sup>5</sup> The Intercontinental Dictionary Series (2023) dictionaries for many of the Northeast Caucasian lects examined here, such as Archi, contain the velar fricatives [x] and [ɣ] where other sources, such as the online dictionary of Archi (Chumakina et al., 2007) contain the uvular fricatives [χ] and [ʁ], respectively. As the treatment is consistent, i.e. all of these sounds in the Intercontinental Dictionary Series are velar where all are uvular in the online Archi dictionary, and the uvular and velar fricatives are treated together in the SCA and LexStat methods used for detection (List, 2012a; List, 2012b), I have elected to utilize the forms as given by the Intercontinental Dictionary Series with the understanding that they may, in fact, be uvular in nature.

For this set, the Avar and Archi items are identical. Given that Archi and Avar diverged long ago, it would be unexpected for an item with the same meaning in both languages to have an identical phonological shape. Additionally, the [cʰ] in Avar corresponds more with [c] or [čʰ] in Archi (Nichols, 2003). Overall, given the unexpected similarity in items, it is most likely that this item is borrowed into Archi from Avar.

Table 57: Set associated with ‘tribe, clan’

Language	Alignment
Avar	a h l u
Archi	a h l u

For this set, the items are once again identical. In addition to the fact that they are identical, there do not seem to be many roots related to this one in the Lezgian language branch other than possibly [el] (Budukh) (*The Intercontinental Dictionary Series*, 2023). Given that the items in Avar and Archi are identical even though the languages themselves are not closely related, it is likely that the Archi form is a borrowing from Avar.

Table 58: Set associated with ‘widow’

Language	Alignment
Avar	qʰ o r o l a y
Archi	qʰ o r o l a y

For this set, the Archi and Avar items are identical. As stated previously, given that Archi and Avar diverged likely thousands of years ago, it would be highly unexpected for the two languages to have an identical item with the same semantic meaning. For this reason, the item in Archi is likely a loanword from Avar.

Table 59: Set associated with ‘yesterday’

Language	Alignment
Avar	s o n - -
Archi	ṣ a n y i

For this item, Chumakina (2009b, p. 444) states that it is a borrowing from the Avar form. The [yi] does not appear to be a separate suffix in Archi, so it is not clear where this portion of the lexical item would come from, or why the change from [s] to [s:] occurred from Avar to Archi. The vowel quality is also different between the two, with [o] appearing in Avar and [a] appearing in Archi. This is not what would be expected given that Archi has a tendency to raise vowels in loanwords from Avar, not lower them (Chumakina, 2009b, p. 440). For these reasons, I am uncertain of the status of the Archi item as a loanword from Avar. More items showing the [s].] contrast that are clearly loanwords or clearly inheritance would be useful. The consonant correspondences in Nichols (2003, p. 250) demonstrate that [ṣ] in Archi would expect to pattern with [c] in Avar, which may lend credence to this item being a loanword at some point in time, since the consonant correspondences also do not match what would be expected given direct inheritance.

#### 4.2.3 Possible Borrowings from Lak

This section covers the results of hypothesis four, indicating loanwords from Lak into Archi. There are ten items to be discussed.

Table 60: Set associated with ‘army’

Language	Alignment
Lak	a <sup>s</sup> r a l
Lak Shali	a <sup>s</sup> r a l

Archi      a<sup>h</sup> r i -

For this set, the Lak and Archi items are similar with the exception of the absence of the second consonant in Archi and the differing quality of the second vowel. Lak has [al] while Archi has [i]. The raising of the vowel from [a] in Lak to [i] in Archi is not unexpected of a loanword, as this can occur in loanwords from Lak as well as Avar, as discussed in many examples above (Chumakina, 2009b, p. 440). The final [l] may have been lost or may be a suffix in Lak that was not borrowed. Overall, given the similarity, it is likely that the Archi form is a loanword from Lak.

Table 61: Set associated with ‘bull’

Language	Alignment
Lak	b u y a
Lak Shali	b u y a
Archi	b u y a

For this item, the Lak and Archi forms are identical, which would be unexpected given the historical divergence of the lects. Additionally, the consonant correspondences given in Nichols (2003, p. 249) indicate that [y] in Lak would be expected to correspond with [q] in Archi given direct inheritance. Therefore, it seems likely that the item in Archi is the result of contact with Lak.

Table 62: Set associated with ‘eyebrow’

Language	Alignment
Lak	i t̄ a - c' a n i
Lak Shali	i t̄ a - c' a n i
Archi	- d a r c' a n -

For this set, Chumakina (2009b, p. 445) states that the item in Archi is a loanword from Lak. It is likely that the Lak item is a compound with [īfa-] meaning ‘eye’ (Starostin, 2011), although it is not clear what the [c’ani] portion would be referring to. Therefore, the [c’an] segment from the Archi item is likely borrowed from Lak, but the [dar] segment may need further investigation.

Table 63: Set associated with ‘frog’

Language	Alignment
Lak	o <sup>s</sup> r w a t’ i
Lak Shali	o <sup>s</sup> r w a t’ i
Archi	u <sup>s</sup> r b i t’ i

For this set, the Archi item and Lak items are extremely similar; the only difference is in the first and second vowels of each, as well as the [w]~[b] for the second consonant. The first vowel of the Lak items is [o<sup>s</sup>] while in Archi it is [u<sup>s</sup>], and the second vowel in the Lak items is [a] while it is [i] in Archi. These alterations would fit with statements in Chumakina (2009b, p. 440) indicating that loanwords from Lak into Archi often undergo vowel-raising. Additionally, given that Lak and Archi are from different branches of the Northeast Caucasian language family, it would be highly unlikely for them to be this similar if the items in both lects were the result of genetic inheritance. As a result, it is likely that the Archi item is the result of borrowing from Lak.

Table 64: Set associated with ‘glove’

Language	Alignment
Lak	k a t’ a
Lak Shali	k a t’ a
Archi	k <sup>w</sup> a t’ i

For this set, the items are similar with the exception of the final vowel, which is [a] in Lak and [i] in Archi, and the initial consonant, which is [k] in Lak and a labialized [k<sup>w</sup>] in Archi. The final vowel can be explained by the tendency for loanwords into Archi from Lak to have raised vowels (Chumakina, 2009b, p. 440). This set is actually given as an example indicating this trend in Chumakina (2009b, p. 440), but her set the Lak form is given as [k<sup>w</sup>at'a] with the same labialized [k<sup>w</sup>] as in Archi. This would make the argument for the Archi form being a loanword from Lak stronger, as there would not otherwise be a clear reason for a [k] to become labialized when borrowed from Lak to Archi.

Table 65: Set associated with 'lake'

Language	Alignment
Lak	b a <sup>s</sup> r -
Lak Shali	b <sup>s</sup> a r -
Archi	b a <sup>s</sup> r i

For this set, the Lak and Archi items are similar with the exception of the final consonant and the placement of the pharyngealization in the Lak Shali item. Standard Lak does not contain pharyngealized consonants but does contain pharyngealized vowels (Friedman, 2020, pp. 204–205), so it is more likely in my opinion that the pharyngealization marker was typed on the wrong side of the vowel in the dictionary entry for Lak Shali than that this lect has pharyngealized [b<sup>s</sup>]. Given that the items are so similar between Lak and Archi, it is most likely that the Archi item is a loanword from Lak.

Table 66: Set associated with ‘owl’

Language	Alignment
Lak	i s u
Lak Shali	i s u
Archi	i s u

For this set, the Lak and Archi items are identical. This is unexpected given that Lak and Archi diverged long ago. Additionally, many other Lezgian forms differ, such as [t’ib] (Lezgian), [t’ib] (Rutul), [t’iḡ] (Southern Tabasaran), and [t’ub] (Kyrts) (*Intercontinental Dictionary Series*, 2003). Given that the Lezgian forms differ from the Archi form and that the Archi form is identical to the Lak form, it is likely that the Archi item is a loanword from Lak.

Table 67: Set associated with ‘root’

Language	Alignment
Lak	m a r $\bar{x}$ a
Lak Shali	m a r $\bar{x}$ a
Archi	m a r $\bar{x}$ u

For this set, The Lak and Archi forms are the same with the exception of the final vowel. While [m] in Lak can correspond with [m] in Archi, [ $\bar{x}$ ] in Lak is expected to correspond with [x] in Archi (Nichols, 2003, pp. 253–254). The final vowel in both lects of Lak is [a] while it is [u] in Archi; Chumakina (2007, p. 440) states that loans from Lak into Archi often undergo vowel-raising, so a change from [a] to [u] would not be



unexpected. Lastly, Lak and Archi diverged long ago, so it would be highly unexpected for items in these lects to be this phonologically similar and have the same meaning. As such, the Archi item is likely a loanword from Lak.

Table 68: Set associated with ‘short’

Language	Alignment
Lak	k u t' a s̄ a -
Lak Shali	k u t' a s̄ a -
Archi	k ū t' a t̄ u t

For this set, the initial portion of each item, [kut'a] in Lak and [kūt'a] in Archi, are similar. In the Lak form the ending [-s̄a] is an attributive marker (Schulze, 2016, p. 3626), and in the Archi form [-t̄u-] is an adjectival derivational suffix (Chumakina, 2009b, p. 441) and the final [-t] is a gender marker for gender IV in the singular form (Chumakina, 2016, p. 3598). As such, only the initial portions of each form are compared. The forms [kut'a] in Lak and [kūt'a] in Archi are nearly identical except for the length of the vowel in Archi. We already know that loanwords from Lak into Archi can have the potential to raise the vowel (Chumakina, 2009b, p. 440); perhaps in this case the vowel lengthened because it was already as high as it could go. Alternatively, perhaps Archi underwent a vowel-lengthening process after the loanword was borrowed. More items would need to be found demonstrating a [ū] in Archi and a [u] in Lak that are loanwords in order to investigate this difference. Overall, given the similarity of the forms [kut'a] in Lak and [kūt'a] in Archi, especially in light of the fact that they diverged a long time ago, it is likely that the Archi form is a loanword from Lak.

Table 69: Set associated with ‘snow’

Language	Alignment
Lak	m a r $\bar{x}$ a l a
Lak Shali	m a r $\bar{x}$ a l a
Archi	m a r $\bar{x}$ a l a

For these forms, the items in Lak and Archi are identical. We would expect a  $[\bar{x}]$  in Lak to correspond with  $[x]$  in Archi given the correspondences in (Nichols, 2003, p. 253); however, these same correspondences do not provide a set with  $[\bar{x}]$  in Archi, so it is possible that  $[\bar{x}]$  in Archi could also correspond with  $[\bar{x}]$  in Lak. However, given the time-depth separating the Lak and Lezgian branches of the Northeast Caucasian family, it is unlikely that these two items would retain an identical phonological form and semantic meaning. As such, the item in Archi is likely a loanword from Lak.

## CHAPTER 5. SUMMARY

### 5.1 Borrowings in Context

Upon reviewing the sets, there is a clear direction in which the majority of the loan occurred: from Avar to Archi, with from Lak to Archi in second place. Very few loans were identified as going from Avar to Lak Shali, and none were identified as going from Archi to Lak Shali. It is possible more loans from Avar into Lak might have been identified if the Avar forms were compared with both standard Lak and Lak Shali instead of just Lak Shali; organizing the distribution such that I looked for similarities in only Lak Shali meant that loan from Avar into Lak lects more broadly would not appear. As this was done intentionally to keep the focus on the cluster of villages of Archib, Chitab, and Shalib, the results should then perhaps be taken to demonstrate that there are fewer loanwords borrowed directly from Avar into only Lak Shali than into Archi, but these results cannot speak to the overall proportion of loanwords into Lak lects more broadly. Additionally, as the Intercontinental Dictionary Series did not have a dictionary for the Chitab lect of Avar, loanwords into Chitab Avar from Archi or Lak could not be examined either. It would be highly unlikely that loanwords on the local level from Archi or Lak Shali would penetrate into broader Avar usage, and as such a village-specific list from Chitab would be needed to determine the lexical borrowing from Archi and Lak Shali that results from this small-scale multilingualism.

While 52 loans examined as a part of this research were internal loans, there were also 54 loans from non-Northeast Caucasian languages. Many other loans from Russian and Persian were also filtered out as a result of the lexical distribution approach; as such, there were likely even more loans from external sources than internal sources. Having

such an even split of loans from languages internal to the Northeast Caucasian language family and external to the Northeast Caucasian family is interesting, given that levels of multilingualism for these external languages historically likely did not approach the levels of multilingualism that existed between the villages themselves. Arabic, for example, which provided 15 loanwords to this sample, was not learned from native speakers or used for spoken communication; instead, it was studied by some in the villages formally for religious purposes (Dobrushina, 2013, p. 381). For loanwords in Archi, Chumakina (2009b, p. 439) states that many loans from Arabic are terms related to time or religion specifically.

Kumyk and Azerbaijani, Turkic languages, were more widely spoken in Archib in the late 1800s and early 1900s than they are now (see figure 11 below for levels of multilingualism in Archib), but the rates of multilingualism in these languages never approached the levels of Lak or Avar. Regardless, there are still many loans from Turkic languages into Archi, particularly in the semantic realms of clothing and grooming, warfare and hunting, and social and political relations (Chumakina, 2009b, p. 438). With regards to Russian, it was not widely spoken in Dagestan until the 1950s when Russian teachers were sent to villages in Dagestan to teach the language in school (Dobrushina, 2013, p. 382), but it has regardless contributed many loanwords into Archi, especially in the semantic fields of law, warfare and hunting, and the house (Chumakina, 2009b, p. 438).

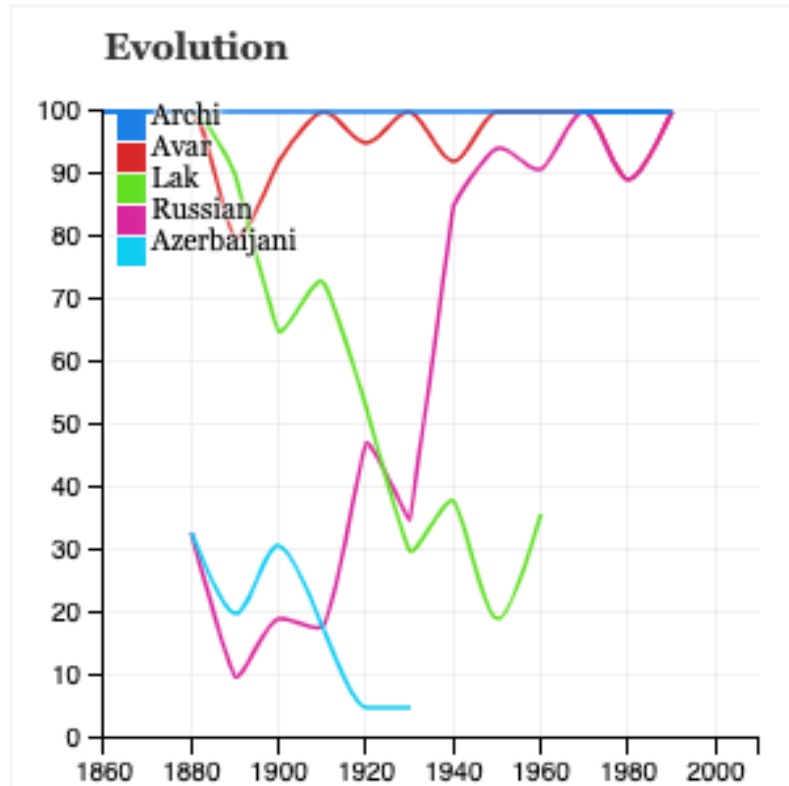


Figure 11: Multilingualism over time in Archib (Dobrushina et al., 2017)

Overall, if one expected the levels of multilingualism to align perfectly with the amount of loanwords, it may be surprising to note some of the variance involved. Russian is now spoken fairly widely and has a large number of loanwords, but Arabic is not spoken and also contributes many loanwords. Avar and Lak were spoken at similar levels of multilingualism as each other in the past, but Avar has also seemed to contribute more loanwords, at least to Archi. This aligns with results found in a previous study of lexical borrowing in Dagestan more broadly, in which more loanwords were borrowed from languages that acted as local lingua francas than from languages used solely to communicate with L1 speakers of that language (Daniel et al., 2021, p. 553). While those in Archib did use Avar to communicate with those in Chitab, Avar has also functioned as

a local lingua franca, perhaps allowing more loanwords to enter the Archi lect in Archib as a result of its use as a form of communication with those other than Avar L1 speakers. In general, the amounts of loanwords from each recipient language into each donor language has the potential to reveal historical factors about the contact situations involved, but one must also keep in mind the socio-historical and socio-cultural factors that have the potential to alter the effects of language contact on the linguistic system; the simple fact that there is a high degree of multilingualism involving a given language does not necessarily ensure a high rate of lexical borrowing, especially compared to other languages involved in the contact situation.

## **5.2 Conclusion**

This work began as an effort to utilize a computational method for cognate detection, the LexStat program within the LingPy library (List, 2012a; List & Forkel, 2021). What was already suspected has become clear as a result; these computational methods are not perfect. While LexStat is designed to detect true cognates (List, 2012a), only 90 out of 283 concepts (31.8%) detected as similar and examined here were true cognates. Many were loans, either from internal sources or various external sources. Looking at the large numbers of both true cognates and loanwords detected, it is clear that deep knowledge of the languages involved as well as language documentation and resources such as those used in the interpretation of the results above (Chumakina, 2009b, 2016; Khalilov & Khalilova, 2016; Nichols, 2003; Schulze, 2016; Starostin, 2011) is still incredibly important for being able to interpret the results of this kind of approach. As a result, I refer to this method as a “computer-assisted” approach in line with Wu et al. (2020) in order to recenter the importance of manually applying the knowledge and methods of

historical linguists. There are also clearly some difficulties with the computational methods for cognate and borrowing detection, including the very idea of a concept list and how to handle synonyms and the treatment of semantic drift.

Future work in this area could include further investigating the loanwords and cognates investigated in this paper for more patterns of consonant correspondences as well as the exact history of the loanwords involved. More concept lists on the village level, such as those developed for the DagSwadesh project (Filatov & Daniel, n.d.), would also be useful in determining actual lexical borrowing between individual villages as opposed to dictionary lists on a larger scale. Overall, however, I believe that in regions such as Dagestan where many languages have existed in close proximity to one another for centuries and additionally where there are not many written records to reveal earlier stages of the development of the languages of the regions, it is important to consider any and as much of the data as possible. This can include lexical data that is high-quality and village-specific, but also includes sociolinguistic information such as that provided in Dobrushina (2013) and village-specific syntactic and typological information as well. Computational methods can help us to pinpoint areas for future investigation, but on their own are not able to replace the knowledge and skills of a trained linguist.

## REFERENCES

- Amigó, E., Gonzalo, J., Artiles, J., & Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4), 461–486. <https://doi.org/10.1007/s10791-008-9066-8>
- Amiridze, N. (2019). Languages of the Caucasus and contact-induced language change. *STUF - Language Typology and Universals*, 72(2), 185–192. <https://doi.org/10.1515/stuf-2019-0007>
- Atkinson, Q. D. (2013). The descent of words. *Proceedings of the National Academy of Sciences*, 110(11), 4159–4160. <https://doi.org/10.1073/pnas.1300397110>
- Belyaev, O. (2014). Annotated Swadesh wordlists for the Dargwa group (North Caucasian family). In G. Starostin (Ed.), *The Global Lexicostatistical Database*. Moscow Higher School of Economics & Santa Fe Institute. <http://starling.rinet.ru/new100/>
- Bergsma, S., & Kondrak, G. (2007). Multilingual cognate identification using intergar linear programming. *Proceedings of the International Workshop on Acquisition and Management of Multilingual Lexicons*, 11–18. [https://acl-bg.org/proceedings/2007/RANLP\\_W6%202007.pdf#page=21](https://acl-bg.org/proceedings/2007/RANLP_W6%202007.pdf#page=21)
- Bouchard-Côté, A., Hall, D., Griffiths, T. L., & Klein, D. (2013). Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11), 4224–4229. <https://doi.org/10.1073/pnas.1204678110>
- Campbell, L. (2013). *Historical Linguistics* (3rd ed.). The MIT Press.
- Chechuro, I. (2021). Lexical convergence reflects complex historical processes: A case study of two borderline regions of Russia. In D. Forker & L. A. Grenoble (Eds.), *IMPACT: Studies in Language, Culture and Society* (Vol. 50). John Benjamins Publishing Company. <https://doi.org/10.1075/impact.50.02che>
- Chechuro, I., Daniel, M., & Verhees, S. (2021). Small-scale multilingualism through the prism of lexical borrowing. *International Journal of Bilingualism*, 25(4), 1019–1039. <https://doi.org/10.1177/13670069211023141>
- Chumakina, M. (2009a). *Archi Vocabulary*. In M. Haspelmath & U. Tadmor (Eds.), *World Loanword Database*. Max Planck Institute for Evolutionary Anthropology. <http://wold.clld.org/vocabulary/16>
- Chumakina, M. (2009b). Loanwords in Archi, a Nakh-Daghestanian language of the North Caucasus. In M. Haspelmath & U. Tadmor (Eds.), *Loanwords in the World's Languages* (pp. 430–446). Walter de Gruyter. <https://doi.org/10.1515/9783110218442.430>
- Chumakina, M. (2016). Archi. In *Volume 5 Word-Formation: An International Handbook of the Languages of Europe* (pp. 3595–3604). De Gruyter Mouton. <https://doi.org/10.1515/9783110226621>
- Chumakina, M. (2020). Archi. In M. Polinsky (Ed.), *The Oxford Handbook of the Languages of the Caucasus* (pp. 281–316). Oxford University Press.
- Chumakina, M., Brown, D., Corbett, G. G., & Quilliam, H. (2007). *A Dictionary of Archi: Archi-Russian-English* (Online edition). University of Surrey.



- Ciobanu, A. M., & Dinu, L. P. (2015). Automatic Discrimination between Cognates and Borrowings. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 431–437. <https://doi.org/10.3115/v1/P15-2071>
- Ciobanu, A. M., & Dinu, L. P. (2020). Automatic Identification and Production of Related Words for Historical Linguistics. *Computational Linguistics*, 45(4), 667–704. [https://doi.org/10.1162/coli\\_a\\_00361](https://doi.org/10.1162/coli_a_00361)
- Comrie, B. (2008). Linguistic Diversity in the Caucasus. *Annual Review of Anthropology*, 37, 131–143.
- Comrie, B., & Khalilov, M. (2009). Bezhta Vocabulary. In M. Haspelmath & U. Tadmor (Eds.), *World Loanword Database*. Max Planck Institute for Evolutionary Anthropology. <http://wold.clld.org/vocabulary/16>
- Daniel, M., Chechuro, I., Verhees, S., & Dobrushina, N. (2021). Lingua francas as lexical donors: Evidence from Daghestan. *Language*, 97(3), 520–560. <https://doi.org/10.1353/lan.2021.0046>
- Dobrushina, N. (2013). How to study multilingualism of the past: Investigating traditional contact situations in Daghestan. *Journal of Sociolinguistics*, 17(3), 376–393. <https://doi.org/10.1111/josl.12041>
- Dobrushina, N. (2023). Language ideology in an endogamous society: The case of Daghestan. *Journal of Sociolinguistics*, 27(2), 159–176.
- Dobrushina, N., Daniel, M., & Koryakov, Y. (2020a). Atlas of multilingualism in Daghestan: A case study in diachronic sociolinguistics. *Languages of the Caucasus*, 4. <https://www.proquest.com/docview/2441548654/abstract/256D0630623340ADP/Q/1>
- Dobrushina, N., Daniel, M., & Koryakov, Y. (2020b). Languages and Sociolinguistics of the Caucasus. In M. Polinsky (Ed.), *The Oxford Handbook of the Languages of the Caucasus* (pp. 26–66). Oxford University Press.
- Dobrushina, N., Kozhukhar, A., & Moroz, G. (2019). Gendered Multilingualism in highland Daghestan: Story of a loss. *Journal of Multilingual and Multicultural Development*, 40(2), 115–132. <https://doi.org/10.1080/01434632.2018.1493113>
- Dobrushina, N., & Kultepeina, O. (2021). The rise of a lingua franca: The case of Russian in Dagestan. *International Journal of Bilingualism*, 25(1), 338–358. <https://doi.org/10.1177/1367006920959717>
- Dobrushina, N., & Moroz, G. (2021). The speakers of minority languages are more multilingual. *International Journal of Bilingualism*, 25(4), 921–938. <https://doi.org/10.1177/13670069211023150>
- Dobrushina, N., Staferova, D., & Belokon, A. (Eds.). (2017). *Atlas of Multilingualism in Dagestan Online*. Linguistic Convergence Laboratory. <https://multidagestan.com>
- Dolgopolsky, A. B. (1986). A Probabilistic Hypothesis Concerning the Oldest Relationships Among the Language Families of Northern Eurasia. In V. V. Shevoroshkin & T. L. Markey (Eds. & Trans.), *Typology Relationship and Time* (pp. 27–50). Karoma Publishers, Inc.

- Epps, P., Huehnergard, J., & Pat-El, N. (2013). Introduction: Contact Among Genetically Related Languages. *Journal of Language Contact*, 6(2), 209–219. <https://doi.org/10.1163/19552629-00602001>
- Facts about Russia's republic of Dagestan | Reuters. (2023, October 29). Reuters. <https://www.reuters.com/world/europe/facts-about-russias-republic-dagestan-2023-10-30/>
- Filatov, K., & Daniel, M. (Eds.). (n.d.). *DagSwadesh: 100 Swadesh lists from Daghestan*. An online database of basic vocabulary divergence across neighbour villages. Moscow: Linguistic Convergence Laboratory, HSE University. [lingconlab.github.io/dagswadesh/index.html](http://lingconlab.github.io/dagswadesh/index.html)
- Forker, D. (2020). Avar. In M. Polinsky (Ed.), *Handbook of Languages of the Caucasus*. Oxford University Press.
- Friedman, V. A. (2020). Lak. In M. Polinsky (Ed.), *The Oxford Handbook of the Languages of the Caucasus* (pp. 201–242). Oxford University Press.
- Ganenkov, D., & Maisak, T. (2020). Nakh-Daghestanian Languages. In M. Polinsky (Ed.), *The Oxford Handbook of the Languages of the Caucasus* (pp. 87–146). Oxford University Press.
- Gigineishvili, B. K. (1977). *Sravnitel' naja fonetikadagestanskix jazykov*. Tbilisi University.
- Hall, D., & Klein, D. (2011). Large-Scale Cognate Recovery. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 344–354. <https://aclanthology.org/D11-1032>
- Hantgan, A., & List, J.-M. (2018). Bangime: Secret Language, Language Isolate, or Language Island? <https://hal.science/hal-01867003>
- Haspelmath, M., & Tadmor, U. (Eds.). (2009). *World Loanword Database*. Max Planck Institute for Evolutionary Anthropology. <http://wold.clld.org>
- Heggarty, P. (2010). Beyond lexicostatistics: How to get more out of 'word list' comparisons. *Diachronica*, 27, 301–324. <https://doi.org/10.1075/dia.27.2.07heg>
- Heggarty, P. (2021). Cognacy Databases and Phylogenetic Research on Indo-European | *Annual Reviews*. *Annual Review of Linguistics*, 7, 371–394.
- Jäger, G. (2019). Computational historical linguistics. *Theoretical Linguistics*, 45(3–4), 151–182. <https://doi.org/10.1515/tl-2019-0011>
- Kassian, A. (2011). Annotated Swadesh wordlists for the Lezgian group (North Caucasian family). In G. Starostin (Ed.), *The Global Lexicostatistical Database*. Moscow Higher School of Economics & Santa Fe Institute. <http://starling.rinet.ru/new100/>
- Kassian, A. (2013). Annotated Swadesh wordlists for the Tsezic group (North Caucasian family). In G. Starostin (Ed.), *The Global Lexicostatistical Database*. Moscow Higher School of Economics & Santa Fe Institute. <http://starling.rinet.ru/new100/>
- Kassian, A. (2015). Towards a Formal Genealogical Classification of the Lezgian Languages (North Caucasus): Testing Various Phylogenetic Methods on Lexical Data. *PLOS ONE*, 10(2), e0116950. <https://doi.org/10.1371/journal.pone.0116950>
- Kassian, A. (2017). Linguistic homoplasy and phylogeny reconstruction. The cases of Lezgian and Tsezic languages (North Caucasus). *Folia Linguistica*, 51(38–1), 217–262. <https://doi.org/10.1515/flih-2017-0008>

- Kassian, A. S., & Testeleets, Y. G. (2017). Pitfalls of shared innovations: Genealogical classification of the Tsezic languages and the controversial position of Hinukh (North Caucasus). *Lingua*, 196, 98–118.  
<https://doi.org/10.1016/j.lingua.2017.06.011>
- Khalilov, M. (2023a). Archi (variety 1) dictionary. In M. R. Key & B. Comrie (Eds.), *The Intercontinental Dictionary Series*. Max Planck Institute for Evolutionary Anthropology. <http://ids.cld.org/contributions/62>
- Khalilov, M. (2023b). Archi (variety 2) dictionary. In M. R. Key & B. Comrie (Eds.), *The Intercontinental Dictionary Series*. Max Planck Institute for Evolutionary Anthropology. <http://ids.cld.org/contributions/62>
- Khalilov, M. (2023c). Avar dictionary. In M. R. Key & B. Comrie (Eds.), *The Intercontinental Dictionary Series*. Max Planck Institute for Evolutionary Anthropology. <http://ids.cld.org/contributions/62>
- Khalilov, M. (2023d). Lak dictionary. In M. R. Key & B. Comrie (Eds.), *The Intercontinental Dictionary Series*. Max Planck Institute for Evolutionary Anthropology. <http://ids.cld.org/contributions/62>
- Khalilov, M. (2023e). Lak (Shali dialect) dictionary. In M. R. Key & B. Comrie (Eds.), *The Intercontinental Dictionary Series*. Max Planck Institute for Evolutionary Anthropology. <http://ids.cld.org/contributions/62>
- Khalilov, M., & Khalilova, Z. (2016). Avar. In *Volume 5 Word-Formation: An International Handbook of the Languages of Europe*. De Gruyter Mouton.  
<https://doi.org/10.1515/9783110226621>
- Kibrik, A. E., & Kodzasov, S. V. (1988). *Sopostavitel'noe izuchenie dagestanskix jazykov: Glagol*. Moscow State University.
- Kibrik, A. E., & Kodzasov, S. V. (1990). *Sopostavitel'noe izuchenie dagestanskix jazykov: Imja. Fonetika*. Moscow State University.
- Kikusawa, R. (2015). The Austronesian Language Family. In C. Bowern & B. Evans (Eds.), *The Routledge handbook of historical linguistics*. Routledge.
- Kondrak, G. (2009). Identification of Cognates and Recurrent Sound Correspondences in Word Lists. *Traitement Automatique Des Langues*, 50(2), 201–235.
- Koryakov, Y. (2020). Maps. In M. Polinsky (Ed.), *The Oxford handbook of languages of the Caucasus* (pp. xxvii–xxx). Oxford University Press.
- List, J.-M. (2012a). LexStat: Automatic Detection of Cognates in Multilingual Wordlists. *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 117–125. <https://aclanthology.org/W12-0216>
- List, J.-M. (2012b). SCA: Phonetic Alignment Based on Sound Classes. In D. Lassiter & M. Slavkovik (Eds.), *New Directions in Logic, Language and Computation* (Vol. 7415, pp. 32–51). Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-642-31467-4\\_3](https://doi.org/10.1007/978-3-642-31467-4_3)
- List, J.-M. (2019). Automated methods for the investigation of language contact, with a focus on lexical borrowing. *Language & Linguistics Compass*, 13(10), N.PAG-N.PAG. <https://doi.org/10.1111/lnc3.12355>
- List, J.-M. (preprint). *Computational Approaches to Historical Language Comparison*. <https://hcommons.org/deposits/item/hc:45473/>

- List, J.-M., & Forkel, R. (2021). LingPy. A Python library for historical linguistics (2.6.9) [Computer software]. Max Planck Institute for Evolutionary Anthropology. <https://lingpy.org>
- List, J.-M., & Forkel, R. (2022). Automated identification of borrowings in multilingual wordlists. *Open Research Europe*, 1(79). <https://doi.org/10.12688/openreseurope.13843.3>
- List, J.-M., Greenhill, S. J., & Gray, R. D. (2017). The Potential of Automatic Word Comparison for Historical Linguistics. *PLOS ONE*, 12(1), e0170046. <https://doi.org/10.1371/journal.pone.0170046>
- McMahon, A., & McMahon, R. (2005). *Language Classification by Numbers*. Oxford University Press.
- Moro, F. R., Sulistyono, Y., & Kaiping, G. A. (2023). Detecting Papuan Loanwords in Alorese: Combining Quantitative and Qualitative Methods. In M. Klammer & F. Moro (Eds.), *Traces of Contact in the Lexicon: Austronesian and Papuan Studies*. BRILL. <https://doi.org/10.1163/9789004529458>
- Nichols, J. (1997). Modeling Ancient Population Structures and Movement in Linguistics. *Annual Review of Anthropology*, 26, 359–384.
- Nichols, J. (2003). The Nakh-Daghestanian Consonant Correspondences. In D. A. Holisky & K. Tuite (Eds.), *Current Issues in Linguistic Theory* (Vol. 246, pp. 207–264). John Benjamins Publishing Company. <https://doi.org/10.1075/cilt.246.14nic>
- Nichols, J. (2013). The vertical archipelago: Adding the third dimension to linguistic geography. In P. Auer, M. Hilpert, A. Stukenbrock, & B. Szmrecsanyi (Eds.), *Space in Language and Linguistics* (pp. 38–60). DE GRUYTER. <https://doi.org/10.1515/9783110312027.38>
- Nikolayev, S. L., & Starostin, S. A. (1994). *A North Caucasian Etymological Dictionary*. Asterisk. <https://starlingdb.org/Texts/caucpref.pdf>
- Oakes, M. P. (2000). Computer Estimation of Vocabulary in a Protolanguage from Word Lists in Four Daughter Languages. *Journal of Quantitative Linguistics*, 7(3), 233–243. <https://doi.org/10.1076/jqul.7.3.233.4105>
- Pakendorf, B., Dobrushina, N., & Khanina, O. (2021). A typology of small-scale multilingualism. *International Journal of Bilingualism*, 25(4), 835–859. <https://doi-org.ezproxy.uky.edu/10.1177/13670069211023137>
- Pat-El, N. (2013). Contact or Inheritance? Criteria for distinguishing internal and external change in genetically related languages. *Journal of Language Contact*, 6(2), 313–328. <https://doi.org/10.1163/19552629-00602006>
- Polinsky, M. (2020). Introduction. In M. Polinsky (Ed.), *The Oxford Handbook of Languages of the Caucasus* (pp. 1–26). Oxford University Press.
- Rama, T. (2015). Automatic cognate identification with gap-weighted string subsequences. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1227–1231. <https://doi.org/10.3115/v1/N15-1130>
- Rhyne, J. (2017). *Quantifying the Comparative Method: Applying Computational Approaches to the Balto-Slavic Question*. The University of Georgia.

- Ross, M., & Durie, M. (1996). Introduction. In M. Durie & M. Ross (Eds.), *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*. Oxford University Press.
- Sankoff, G. (2004). Linguistic Outcomes of Language Contact. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The Handbook of Language Variation and Change* (1st ed., pp. 638–668). Wiley.  
<https://doi.org/10.1002/9780470756591.ch25>
- Schulze, W. (2013). Historische und areale Aspekte der Bodenschatz- Terminologie in den ostkaukasischen Sprachen. *Iran and the Caucasus*, 17(3), 295–320.  
<https://doi.org/10.1163/1573384X-20130305>
- Schulze, W. (2016). Lak. In *Volume 5 Word-Formation: An International Handbook of the Languages of Europe* (pp. 3622–3637). De Gruyter Mouton.  
<https://doi.org/10.1515/9783110226621>
- Schulze, W. (2017). The Comparative Method in Caucasian linguistics. In J. Klein, B. Joseph, & M. Fritz (Eds.), *Handbook of Comparative and Historical Indo-European Linguistics* (pp. 105–114). De Gruyter.  
<https://doi.org/10.1515/9783110261288-011>
- Sokal, R. R., & Michener, C. D. (1958). A Statistical Method for Evaluating Systematic Relationships. *University of Kansas Science Bulletin*, 38(22).  
[https://ia800703.us.archive.org/5/items/cbarchive\\_33927\\_astatisticalmethodforevaluatin1902/astatisticalmethodforevaluatin1902.pdf](https://ia800703.us.archive.org/5/items/cbarchive_33927_astatisticalmethodforevaluatin1902/astatisticalmethodforevaluatin1902.pdf)
- St Arnaud, A., Beck, D., & Kondrak, G. (2017). Identifying Cognate Sets Across Dictionaries of Related Languages. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2519–2528.  
<https://doi.org/10.18653/v1/D17-1267>
- Starostin, G. (Ed.). (2011). *The Global Lexicostatistical Database*. Moscow: Higher School of Economics, & Santa Fe: Santa Fe Institute.  
<http://starling.rinet.ru/new100/>
- Steiner, L., Cysouw, M., & Stadler, P. (2011). A Pipeline for Computational Historical Linguistics. *Language Dynamics and Change*, 1(1), 89–127.  
<https://doi.org/10.1163/221058211X570358>
- Swadesh, M. (1952). Lexico-Statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society*, 96(4), 452–463.
- Tadmor, U., Haspelmath, M., & Taylor, B. (2010). Borrowability and the Notion of Basic Vocabulary. *Diachronica*, 27, 226–246. <https://doi.org/10.1075/dia.27.2.04tad>
- The Intercontinental Dictionary Series. (2023). Max Planck Institute for Evolutionary Anthropology. <https://ids.clld.org>
- Thomason, S. G. (2001). *Language Contact: An Introduction*. Georgetown University Press.
- Thomason, S. G., & Kaufman, T. (1988). *Language Contact, Creolization, and Genetic Linguistics*. University of California Press.
- Trudgill, P. (2010). Contact and Sociolinguistic Typology. In R. Hickey (Ed.), *The Handbook of Language Contact* (1st ed., pp. 299–319). Wiley.  
<https://doi.org/10.1002/9781444318159.ch15>

- Trudgill, P. (2015). Societies of Intimate and Linguistic Complexity. In *Language Structure and Environment: Social, cultural, and natural factors* (pp. 133–148).
- Tuite, K. (1999). The myth of the Caucasian Sprachbund: The case of ergativity. *Lingua*, 108(1), 1–29. [https://doi.org/10.1016/S0024-3841\(98\)00037-0](https://doi.org/10.1016/S0024-3841(98)00037-0)
- Urban, M. (2020). Mountain linguistics. *Language and Linguistics Compass*, 14(9), e12393. <https://doi.org/10.1111/lnc3.12393>
- Winford, D. (2003). *An Introduction to Contact Linguistics*. Blackwell Publishing.
- Wixman, R. (1980). *Language aspects of ethnic patterns and processes in the north Caucasus*. University of Chicago Press.
- Wu, M.-S., Schweikhard, N., Bodt, T., Hill, N., & List, J.-M. (2020). Computer-Assisted Language Comparison: State of the Art. *Journal of Open Humanities Data*, 6, 2. <https://doi.org/10.5334/johd.12>
- Zaitsev, K., & Minchenko, A. (2022). Automatic Detection of Borrowings in Low-Resource Languages of the Caucasus: Andic branch. *Proceedings of the First Workshop on NLP Applications to Field Linguistics*, 34–41. <https://aclanthology.org/2022.fieldmatters-1.4>

## VITA

**Bonnie Eleanor Wren-Hardin** was raised in New Jersey and earned a Bachelor of Arts in Linguistics with a minor in German from the University of Oklahoma, followed by a Professional Master of Arts in Teaching English to Speakers of Other Languages (TESOL) from the University of Oklahoma. As a graduate student at the University of Kentucky, she taught as instructor of record for the Department of Writing, Rhetoric, and Digital Studies and as a tutor in the Robert E. Hemenway Writing Center. Upon publication of this thesis, she will have completed a Master of Arts in Linguistic Theory and Typology from the University of Kentucky.