Manana Bukia [1]

Ivane Javakhishvili Tbilisi State University, Georgia

Rati Skhirtladze

Muskhelishvili Institute of Computational Mathematics of Georgian Technical University
Caucasus University, Georgia

Comparative Analysis of the Phonemic Structures of Georgian and Abkhazian Languages Using Computational Linguistic Methods[2]

ABSTRACT

This study presents a comparative analysis of the phonemic structures of Georgian and Abkhazian languages using computational linguistic methods and corpus-based methodologies. Employing the diasystemic approach, we analyze large-scale corpora (Georgian: 953 million tokens; Abkhazian: 19.4 million tokens) to reveal systematic patterns in vowel distribution. Our findings confirm the principle of markedness theory in both languages, demonstrating universal dominance of the vowel *a* and the marked status of labial vowels. The study identifies both common typological characteristics (preference for open vowels in anlaut position, labial/non-labial opposition) and distinctive features (six-vowel system in Abkhazian versus five-vowel system in Georgian, specific role of the neutral vowel *ə* in Abkhazian). These data provide an empirical foundation for historical-comparative analysis and contribute to the development of Caucasian areal linguistics.

Keywords: *Georgian language, Abkhazian language, phonemic structure, diasystemic analysis, corpus linguistics, vocalism, computational linguistics.*

Introduction

Languages operate through various interacting levels that simultaneously influence the global linguistic system. Phonology occupies a unique position in this hierarchy: while individual phonemes carry no inherent meaning or linguistic value, they present distinct challenges from other linguistic levels (morphology, syntax). Phonemes combine to form morphemes, words, and sentences, yet their syntagmatic relations differ fundamentally within morphemic and lexical frameworks, necessitating specialized investigation.

The establishment of phonotactic rules throughout a language system benefits from the diasystemic method, which draws upon data from linguistic subsystems (Uturgaidze, 1976). Rules derived through diasystemic methodology prove reliable as they are grounded in properties shared

[1] Corresponding author: manana Bukia, manana.bukia@tsu.ge

[2] „This work was supported by Shota Rustaveli National Science Foundation of Georgia (SRNSFG) [FR-24-5970, Theoretical bases of speech synthesis of Abkhazian language]“

across subsystems.

This article employs the diasystemic method to analyze the Abkhazian phonemic system and compare these findings with Georgian phonemic structures. The novelty of this research lies in applying the diasystemic approach to study the phonemic structures of Georgian and Abkhazian languages, revealing synchronic differences and processes. Based on these results, we attempt a historical-comparative analysis at the diachronic level, advancing our understanding of historical connections between Georgian and Abkhaz-Adyghe languages.

The study encompasses three primary stages:

1. Analysis of vocalism in Georgian and Abkhazian languages;
2. Examination of consonant systems in Georgian and Abkhazian languages;
3. Historical-comparative analysis of Georgian and Abkhaz-Adyghe phonemic structures.

This article focuses on the first stage: investigating vocalism in Georgian and Abkhazian languages utilizing computational linguistic methods. Through frequency analysis of language corpora, we reveal systemic regularities that would be impractical to identify through traditional methods.

Literature Review

Trubetzkoy (1931) pioneered the comparison of accents in their synchronous states rather than through historical development. He categorized sound differences between dialects into three types, with particular emphasis on phonological inventory and contextual constraints. His interest centered on the phonetic realization of phonemes across different linguistic subsystems.

Building upon Trubetzkoy's work, Weinreich (1954) proposed synthesizing linguistic geography with descriptive linguistics by applying structuralist grammar concepts to describe regular correspondences between varieties. He termed this higher-order system a "diasystem," designed to be compatible with individual grammars of all constituent subsystems. In a diasystem, units of analysis represent higher-order abstractions than those in individual systems—just as phonemes in a single variety group into abstract phonemes, phonemes across varieties can be grouped into even more abstract diaphonemes.

Weinreich acknowledged challenges in constructing diasystems, particularly regarding phonemic merging and splitting with divergent consequences between dialects or linguistic subsystems. Following Trubetzkoy (1931), he noted that differences in phonological inventory and etymological distribution could complicate diasystem construction (Weinreich, 1954).

Subsequent researchers expanded this framework. Moulton (1960) identified divergent cases

developing independently in his analysis of Swiss German dialects from Lucerne and Appenzell. Despite both dialects possessing identical sets of eleven short vowel phonemes, only one pair (/i/ ~ /ɪ/) shared a common ancestral vowel from early German. The remaining phonetic similarities resulted from accidental convergence through multiple independent mergers and splits.

Georgian language phonemic analysis has been addressed extensively. Uturgaidze (1976) produced a seminal work applying diasystemic methodology to Georgian phonetics and phonotactic rules, analyzing various classifications by Georgian and international researchers (Zhghenti, Akhvlediani, Vogt, Robins and Waterson, Shanidze, Chikobava). This systematic approach yielded a comprehensive framework that remains fundamental for phonetics research.

Abkhazian phonemic system studies have employed traditional approaches. Research on Abkhaz-Abaza phonetics-phonology has been conducted by Lomtatidze (1976), Genko (1955), Chirikba (1996), Kuipers (1955), Spruit (1986), Uslar (1862), Marr (1912), Akhvlediani (1949), Rogava (1985), Hewitt (2010), Trubetzkoy (1960), Allen (1956), and others. While these scholars have described phonemic systems and phonotactic rules, research conducted through varying methodologies cannot provide accurate systematic results for comparative analysis, particularly regarding diachronic questions of language kinship within the "Ibero-Caucasian" hypothesis.

This study presents an attempt to elaborate the Abkhazian vowel system using diasystemic methodology, enabling more rigorous comparison with Georgian.

Methodology

A diasystem represents a linguistic analysis framework designed to encode or represent related variants in ways that reflect their structural differences. The integration of computational linguistic methods into diasystemic analysis proves crucial. Computational linguistics enables more effective results and objective data in recording and comparing syntagmatic and paradigmatic language relations. Precise statistical analysis of millions of word forms—including sound frequency, positional distribution (anlaut, inlaut, auslaut), identification of language-specific harmonic complexes, and registration of phonotactic constraints—would be practically impossible without modern computational technologies.

Discussion

Corpus Parameters

For this study, we created large-scale corpora of Georgian and Abkhazian languages:

- Georgian Corpus: 953,008,532 tokens (2,873,730 unique word forms)
- Abkhazian Corpus: 19,417,316 tokens (1,217,820 unique word forms)

The asymmetry between corpora reflects differences in available digital resources for both languages. Nevertheless, both corpora provide statistically representative samples for phonotactic analysis.

Quantitative Analysis of Phoneme Distribution

To analyze positional vowel distribution, we employed the following computational algorithm:

1. Tokenization and Segmentation: Division of corpus into lexical units;
2. Phoneme Annotation: Identification of phonemes corresponding to each grapheme;
3. Positional Classification:
    - Anlaut (initial) – first phoneme in a word;
    - Inlaut (medial) – all phonemes between first and last;
    - Auslaut (final) – last phoneme in a word.

This computational approach represents a systematic attempt to integrate modern corpus linguistic methods into the analysis and comparison of language structures.

Results

Sounds differentiate based on their capacity to distinguish meanings through language's communicative function. This meaning-distinguishing capacity establishes their value as linguistic signs and represents their primary function. Such meaningful distinguishing units hold particular interest for linguists. To differentiate meaning-bearing sounds from non-functional ones, linguists introduced the term "phoneme" in the 1920s-1930s, establishing phonology as a specialized branch studying these units.

Phonology investigates the functioning of speech sounds in language, focusing on sound function rather than phonetics' emphasis on articulation and acoustics. Studying vocalism in Georgian and Abkhazian languages involves determining paradigmatic vowel status and revealing syntagmatic regularities.

Vocalism of Georgian Language

Georgian vowels are traditionally characterized by place of articulation, height, and labialization (Akhvlediani, 1949; Vogt, 1961; Shanidze, 1973; Aronson, 1982; Uturgaidze, 1976). Literary Georgian contains five vowels, analyzed from both articulatory-physical and phonemic perspectives. From an articulatory viewpoint, *e* and *i* are front vowels, while *o* and *u* are back vowels. The vowel *a* is variously classified as central or back by different researchers.

Multiple systematic classifications exist for Georgian vowel production:

1. *i, e, a* as front vowels; *o, u* as back vowels
2. *i, e* as front vowels; *a, o, u* as back vowels
3. *i, e* as front vowels; *a* as central; *o, u* as back vowels
4. *i* as front; *e, a, o* as central; *u* as back

Regarding height, *i* and *u* are high, *e* and *o* are mid, and *a* is low.

Robins and Waterson (1952) observed that *l* shows one variant before vowels *i, e* and another before *a, o, u*. This provides phonological justification for grouping *a* with back vowels, establishing a linguistic and physiological criterion for classification.

Uturgaidze (1976) argues that two features suffice for paradigmatic description of Georgian vowels: labiality and height. This creates:

Labial series: *o, u*

Non-labial series: *a, e, i*

By height:

Low: *a*

Mid: *e, o*

High: *i, u*

The labial correlation in Georgian is supported by frequency data. Marked features (labiality) correlate with lower frequency:

- Non-labial (a, e, i): 78.19%
- Labial (o, u): 21.81%
- Ratio: 3.58:1

These data are confirmed by modern corpus research:

Table 1. Absolute and relative frequencies of Georgian vowels

Vowel	Total	Unique	Anlaut (abs.)	Inlaut (abs.)	Auslaut (abs.)	Anlaut (%)	Inlaut (%)	Auslaut (%)
ა (a)	284,210,119	837,686	40,151,199	202,135,525	41,923,395	47.78	26.91	35.59
ე (e)	209,485,327	660,146	10,893,612	186,050,453	12,541,262	12.97	24.77	10.65
ი (i)	251,481,920	699,389	16,702,030	186,568,567	48,211,323	19.88	24.84	40.94
ო (o)	131,571,636	428,796	5,859,727	112,992,112	12,719,797	6.98	15.04	10.80
უ (u)	76,259,530	247,713	10,404,192	63,460,166	2,395,172	12.39	8.45	2.03
Total	953,008,532	2,873,730	84,010,760	751,206,823	117,790,949	100.00	100.00	100.00

Analysis of Georgian vowel distribution reveals:

1. Frequency hierarchy: a (29.83%) > i (26.38%) > e (21.98%) > o (13.81%) > u (8.00%)
2. Positional distribution patterns:
    - Anlaut: dominance of *a* (47.78%), indicating preference for open-syllable word beginnings;
    - Inlaut: nearly equal distribution of *a, e, i* (24-27%);
    - Auslaut: high concentration of *i* (40.94%), related to Georgian verbal morphology.

Vocalism of Abkhazian Language

Modern Abkhazian contains six vowels. Two (*a* and *ə*) are considered basic (Lomtatidze, 1976; Kuipers, 1955; Trubetzkoy, 1960). Uslar (1862) considered *a, i,* and *u* basic in Abkhazian. Some researchers do not consider *e, i, o, u* as phonemes, arguing they derive from combinations of *a* and *ə* with semivowels *j* and *w* (Lomtatidze, 1976). However, Gvantseladze (2011) argues that all six sounds function as phonemes with word- and form-distinguishing capabilities.

The vowel *a* is a narrow, low vowel, less open than Georgian *a* and produced more frontally. It represents the primary and most frequent vowel, occurring 7,470,873 times in our corpus—twice the frequency of *ə*.

The neutral vowel *ə* is narrower than *a*. It rarely occurs initially, appearing primarily in medial position and finally in stressed form. According to Lomtatidze, initial *ə* is absent, with rare exceptions being recent dialectal variants.

The vowel *e* develops from diphthongs *aj, ja* through intermediate *ej, je* stages via partial assimilation. The vowel *o* derives positionally from diphthongs *aw, wa* through *ow, wo* stages.

The vowels *i* and *u* are positionally derived from diphthongs *əj, jə* and *əw, wə* respectively.

The Abkhazian labial correlation shows:

- Non-labial (a, e, i, ə): 82.97%
- Labial (o, u): 17.03%
- Ratio: 4.87:1

Frequency distribution of Abkhazian vowels by position:

Table 2. Absolute and relative frequencies of Abkhazian vowels

Vowel	Total	Unique	Anlaut (abs.)	Inlaut (abs.)	Auslaut (abs.)	Anlaut (%)	Inlaut (%)	Auslaut (%)
ა (a)	7,470,873	391,184	2,748,746	3,764,928	957,199	47.85	35.70	30.61

ე (e)	1,494,060	110,740	293,998	1,107,298	92,764	5.12	10.50	2.97
ი (i)	3,570,173	244,633	1,652,799	1,317,341	600,033	28.77	12.49	19.19
ო (o)	1,406,409	113,607	150,038	1,044,760	211,621	2.61	9.91	6.77
უ (u)	1,900,624	116,102	646,495	1,015,160	238,969	11.26	9.63	7.64
ы (ə)	3,575,177	241,491	252,919	2,295,362	1,026,903	4.40	21.77	32.84
Total	19,417,316	1,217,820	5,744,978	10,544,849	3,127,489	100.00	100.00	100.00

Analysis reveals:

1. Frequency hierarchy: a (38.48%) > ə (18.41%) > i (18.39%) > u (9.79%) > e (7.69%) > o (7.24%)
2. Positional distribution:
    - Anlaut: dominance of *a* (47.85%), high frequency of *i* (28.77%)
    - Inlaut: dominant *a* (35.70%), high concentration of *ə* (21.77%)
    - Auslaut: maximum representation of *ə* (32.84%)

Conclusion

Corpus linguistic analysis reveals both common typological characteristics and specific features distinguishing Georgian and Abkhazian vowel systems. Markedness theory principles are confirmed in both languages.

Common features include:

1. Universal dominance of vowel *a* in both systems;
2. Marked status of labial/non-labial opposition;
3. Preference for open vowels in anlaut position.

Distinctive features include:

1. Six-vowel system in Abkhazian versus five-vowel system in Georgian
2. Specific functional role of neutral vowel *ə* in Abkhazian
3. Different patterns of positional distribution

These findings provide crucial empirical data for historical-comparative analysis of both languages and contribute to advancing Caucasian areal linguistics. The diachronic implications suggest possible original two-vowel systems in both languages, though this hypothesis requires further syntagmatic frequency analysis for Georgian. The integration of computational linguistic methods opens new avenues for systematic investigation of phonemic structures across Caucasian languages.

References

Akhvlediani, G. (1949). ზოგადი ფონეტიკის საფუძვლები [Foundations of General Phonetics]. Tbilisi: Tbilisi University Press.

Allen, W. S. (1956). Structure and system in the Abaza verbal complex. Transactions of the Philological Society, 55(1), 127-176.

Allen, W. S. (1965). On one-vowel systems. Lingua, 13, 111-124.

Aronson, H. I. (1982). Georgian: A Reading Grammar. Chicago: University of Chicago Press.

Chirikba, V. (1996). Common West Caucasian: The Reconstruction of Its Phonological System and Parts of Its Lexicon and Morphology. Leiden: CNWS Publications.

Chukhua, M. (2017). ქართულ-ჩერქეზულ-აფხაზური ეტიმოლოგიური ძიებანი [Georgian-Circassian-Abkhazian Etymological Studies]. Tbilisi: Saari Publishing.

Genko, A. (1955). Абазинский язык: Грамматический очерк наречия Тапанта [The Abaza Language: A Grammatical Sketch of the Tapanta Dialect]. Moscow: USSR Academy of Sciences.

Gvantseladze, T. (2010). წინარე აფხაზურ-აბაზური ენა: ხმოვანთა სისტემის დიაქრონია [Proto-Abkhaz-Abaza Language: Diachrony of the Vowel System]. In Linguistic Problems of Kartvelology and Abkhazology (Vol. 2). Tbilisi.

Gvantseladze, T. (2011). აფხაზური ენა: სტრუქტურა, ისტორია, ფუნქციონირება [The Abkhazian Language: Structure, History, Functioning]. Tbilisi: Universal Publishing.

Hewitt, G. (2010). Abkhaz: A Comprehensive Self-Tutor. München: Lincom Europa.

Kuipers, A. (1955). The North-West Caucasian languages. Analecta Slavica, 1, 193-206.

Kurdiani, M. (2007). იბერიულ-კავკასიური ენათმეცნიერების საფუძვლები [Foundations of Ibero-Caucasian Linguistics]. Tbilisi: TSU Press.

Lomtatidze, K. (1976). აფხაზური და აბაზური ენების ისტორიულ-შედარებითი ანალიზი I: ფონოლოგიური სისტემა და ფონეტიკური პროცესები [Historical-Comparative Analysis of Abkhazian and Abaza Languages I: Phonological System and Phonetic Processes]. Tbilisi: Metsniereba.

Machavariani, N. (2020). აფხაზური ენის აფიქსები და მოდალური ელემენტები გრამატიკული მიმოხილვითურთ [Affixes and Modal Elements of Abkhazian with Grammatical Overview]. Tbilisi: TSU Press.

Marr, N. (1912). К вопросу о положении абхазского языка среди яфетических [On the Position of Abkhazian among Japhetic Languages]. St. Petersburg: Imperial Academy of Sciences.

Moulton, W. G. (1960). The short vowel systems of Northern Switzerland: A study in structural dialectology. Word, 16, 155-182.

Robins, R. H., & Waterson, N. (1952). Notes on the phonetics of the Georgian word. Bulletin of the School of Oriental and African Studies, 14(1), 55-72.

Rogava, G. (1985). К вопросу о взаимоотношении между сложной системой консонантизма и простой системой вокализма в абхазско-адыгских языках [On the Relationship between Complex Consonantism and Simple Vocalism in Abkhaz-Adyghe Languages]. Annual of Ibero-Caucasian Linguistics, 12, 182-186.

Shanidze, A. (1973). ქართული გრამატიკის საფუძვლები [Fundamentals of Georgian Grammar]. Tbilisi: Tbilisi University Press.

Shengelia, V. (2006). ქართველურ და ჩერქეზულ ენობრივ სისტემათა ისტორიის ზოგი საკითხი [Some Issues in the History of Kartvelian and Circassian Language Systems]. Georgian Language, Tbilisi.

Spruit, A. (1986). Abkhaz Studies. Leiden: University of Leiden.

Trubetzkoy, N. S. (1931). Phonologie et géographie linguistique. Travaux du Cercle Linguistique de Prague, 4, 228-234.

Trubetzkoy, N. S. (1960). Principles of Phonology. Berkeley: University of California Press.

Uslar, P. K. (1862). Этнография Кавказа. Языкознание. Абхазский язык [Ethnography of the Caucasus. Linguistics. The Abkhazian Language]. St. Petersburg: Imperial Academy of Sciences.

Uturgaidze, T. (1976). ქართული ენის ფონემატური სტრუქტურა [The Phonemic Structure of Georgian]. Tbilisi: Metsniereba.

Vogt, H. (1961). ქართული ენის ფონემატური სტრუქტურა [The Phonematic Structure of Georgian]. Tbilisi.

Weinreich, U. (1954). Is a structural dialectology possible? Word, 10(2-3), 388-400.

Zhghenti, S. (1956). ქართული ენის ფონეტიკა [Georgian Phonetics]. Tbilisi: TSU Press.