

Code-Mixing Patterns in a Corpus-Based Megrelian-English Dictionary

*Rusudan Gersamia,
Irina Lobzhanidze
Ilia State University
rgersamia@iliauni.edu.ge,
irina_lobzhanidze@iliauni.edu.ge*

Abstract

This paper examines code-mixing patterns in a corpus-based Megrelian-English dictionary. Megrelian, an endangered Kartvelian language, is characterized by intensive contact with Georgian and, to a lesser extent, Russian. Using data from a morphologically annotated corpus of spoken Megrelian, the study analyzes the structural and functional properties of code-mixing, including insertion, alternation and lexicalized mixing.

The findings show that code-mixing is a systematic and productive phenomenon rather than random interference. Mixed forms are integrated into Megrelian morphosyntax and reflect both linguistic constraints and sociolinguistic factors such as age, region and communicative context.

The paper also demonstrates how such forms can be represented in a corpus-based bilingual dictionary using L2 tagging. This approach contributes to lexicographic practice and provides a more accurate representation of contemporary language use in endangered language documentation.

Keywords: Megrelian; code-mixing; language contact; corpus-based lexicography; endangered language documentation

1 Introduction

Megrelian (ISO 639-3: xmf) is an unwritten Kartvelian language spoken primarily in western Georgia, particularly in the Samegrelo-Zemo Svaneti region, which includes the municipalities of Zugdidi, Senaki, Martvili, Abasha, Chkhorotsku, Khobi and Tsalenjikha. It is also spoken by internally displaced populations from Abkhazia, as well as by Georgian communities residing in the Gali district of the occupied territory. According to UNESCO (2010), Megrelian is classified as an “increasingly endangered” language.

The sociolinguistic environment of Megrelian is influenced by a range of factors. These include political conditions (such as displacement from occupied territories), economic and

industrial influences (for example, the port city of Poti, which has long been a site of intensive language contact) and the impact of neighboring dialects (e.g., the influence of Imeretian dialect of Georgian in Abasha and Gurian dialect - in Poti). In addition, the dominance of Standard Georgian as the state language used in education, media and public life, plays a crucial role in language use (Makharoblidze & Gersamia 2022). Until the late 20th century, Russian also functioned as a major contact language.

As a result, Megrelian speakers are typically bilingual. Standard Georgian, as the official language, is an integral part of their daily lives. And speakers frequently mix languages in everyday communication, producing hybrid expressions and a wide range of mixed lexical forms. The contact between Megrelian and Georgian, and to a lesser extent Russian, provides a clear example of bilingual and multilingual interaction, making Megrelian an important case for the study of code-mixing and language contact.

2 Aim and Methods

The aim of this study is to examine how the outcomes of language contact are reflected in the speech of contemporary Megrelian speakers. In particular, it investigates the linguistic strategies employed in code-mixing, code-switching and borrowing.

The analysis is based on morphologically annotated texts from the Megrelian Language Corpus (MLC) (Gersamia & Lobzhanidze 2021, 2022-2025) as well as data from a corpus-based Megrelian-English dictionary available at <https://xmf.iliauni.edu.ge/>. The corpus consists of spoken language materials, including recorded dialogues, narratives and interviews collected through fieldwork conducted between 2021 and 2024 in urban and rural areas of Samegrelo¹.

The fieldwork followed established principles of language documentation (Austin 2006; Bown 2008). The dataset includes 58 hours of audio-video recordings from 83 speakers, with balanced representation in terms of dialect, age and gender. The material is thematically diverse, covering topics such as traditional rituals, cuisine, oral traditions, folktales, personal narratives etc.

The corpus has been processed in the Fieldwork Language Explorer (FLEX 2024) environment and includes 150 annotated texts, comprising 97 702 tokens and 13 648 lexical and morphological units (roots and affixes), distributed across 9 255 sentences.

The analysis of language contact phenomena combines quantitative and qualitative approaches. First, non-Megrelian elements in the corpus (including roots, affixes, and lexemes) were identified and assigned an L2 tag in the FLEX database. These tagged elements were then extracted and systematically filtered to determine their frequency and distribution.

The results show that L2 elements primarily originate from Georgian (in large numbers) and, to a lesser extent, Russian, alongside a considerable number of internationalisms. Out

1 Grant No FR-21-993, Annotated Corpus of Megrelian Language with Sketch Grammar and Online Dictionary

of 13 648 unique units, 3 593 were marked as L2, including 1 150 fully lexicalized word forms that are not morphologically analyzable.

From an analytical perspective, the study examines:

- the structural integration of L2 elements into Megrelian morphosyntax;
- the types of code-mixing and switching strategies employed by speakers, and;
- the degree of lexicalization of borrowed and hybrid forms.

The analysis of L2-marked items and their integration into a lexicographic resource such as the Megrelian-English dictionary demonstrates that code-mixing is not limited to the use of loanwords or calques. Rather, it constitutes a dynamic and systematic process governed by its own lexical and grammatical principles.

Specifically, elements from the source language (roots, stems, and occasionally affixes) are incorporated into the morphostructure of the recipient language, adapting to its grammatical system.

Consequently, the identification and description of language contact phenomena form an essential component of corpus-based lexicography. Representing such mixed and hybrid forms in the Megrelian-English dictionary and in the lexicon of Megrelian morphemes allows for a more accurate and comprehensive account of contemporary language use.

3 Theoretical Framework

Code-switching is commonly defined as the systematic use of two or more languages within a single discourse (Weinreich 1953; Gumperz 1982 and others). In broad usage, the term encompasses both intersentential switching (between clauses or sentences) and intrasentential switching (within a clause or word), the latter often referred to as code-mixing. Importantly, such switching is not random, but rule-governed, reflecting both grammatical constraints and sociolinguistic conditioning (Poplack 1980 and others).

In this study, we distinguish between code-switching and code-mixing in line with common practice. Code-mixing refers specifically to the use of elements from two or more languages within a single sentence, where lexical and grammatical features from different linguistic systems co-occur (Myers-Scotton 1998; Redouane 2005 and others). By contrast, code-switching involves alternation between languages across sentence boundaries.

A central concept in the analysis is the distinction between the matrix language (L1) and the embedded language (L2) (Myers-Scotton 1993; Auer & Muhamedova 2005). The matrix language provides the grammatical framework of the utterance, while elements from the embedded language are inserted into this structure. Code-mixing may occur at multiple linguistic levels, including grammatical, phonological, lexical, and orthographic (Hoo 2007).

In intrasentential mixing, embedded language elements are typically adapted to the morphosyntactic system of the matrix language. In contrast, in alternational code-switching, speakers shift between languages at structural boundaries, producing sequences that maintain the grammatical integrity of each language.

Borrowing is closely related to these phenomena and is often treated as a special case of code-mixing or code-switching. Borrowed items are typically nativized into the recipient language, showing phonological and morphological integration and often spreading beyond bilingual speakers (Haugen 1956; Poplack et al. 1988). Over time, such forms may become fully lexicalized loanwords, filling lexical or cultural gaps and gradually losing their perception as foreign elements.

For the purposes of analysis, this study adopts a widely used typology of contact-induced phenomena based on Poplack (1980) and subsequent research. Three analytically relevant types are distinguished:

- Insertional mixing i.e. the embedding of words or phrases from a donor language within a Megrelian grammatical frame;
- Alternational switching i.e. the alternation between languages at clause or sentence boundaries;
- Lexicalized mixing i.e. the formation of hybrid structures (e.g., compounds or affixed forms) and the use of shared morphosyntactic patterns across languages.

These categories provide the analytical framework for examining the corpus data and are directly applicable to the representation of mixed forms in the Megrelian-English dictionary.

4 L2 Elements in the Megrelian–English Online Dictionary

4.1 Lexical Composition and Common Domains

The analysis of L2 elements in the Megrelian-English online dictionary reveals that code-mixing is particularly frequent in semantic domains associated with modern life, including technology, education, administration and science. In these domains, lexical items are predominantly internationalisms, which have entered Megrelian primarily through Georgian and typically lack native equivalents (See Table 1).

<i>administ'ratsia</i> 'administration'	<i>globalizatsia</i> 'globalization'	<i>eksp'editsia</i> 'expedition'	<i>k'ompōrt'uli</i> 'comfortable'
<i>ak'ademia</i> 'academy'	<i>ginēk'ōlgia</i> 'gynecology'	<i>eksp'ōnat'i</i> 'exhibit'	<i>k'ōment'ari</i> 'comment'
<i>avt'ōrit'et'i</i> 'authority'	<i>dek'ōratsia</i> 'decoration'	<i>et'imōlgia</i> 'etymology'	<i>k'ōmunik'atsia</i> 'communication'
<i>ak'vat'ōria</i> 'aquatorium'	<i>dēmōsnt'ratsia</i> 'demonstration'	<i>versia</i> 'version'	<i>k'ōmp'iut'eruli</i> 'computer-related'
<i>alergia</i> 'allergy'	<i>dikt'at'ura</i> 'dictatorship'	<i>indust'ria</i> 'industry'	<i>k'ōmersant'i</i> 'merchant'

<i>alumini</i> 'aluminium'	<i>elekt'ropik'atsia</i> 'electrification'	<i>inst'it'ut'i</i> 'institute'	<i>k'oordinat'i</i> 'coordinate'
<i>biznesi</i> 'business'	<i>eksp'eriment'i</i> 'experiment'	<i>inžineri</i> 'engineer'	<i>k'onk'urent'i</i> 'competitor'
<i>biznesmeni</i> 'businessman'	<i>eksp'ansia</i> 'expansion'	<i>k'limat'uri</i> 'climatic'	<i>k'oordintsia</i> 'coordination' etc.

Table 1: International Lexical Items (L2) in Megrelian

4.2 Russian-Origin Vocabulary (Russizms)

Russian-origin lexical items (russizms) are significantly less frequent, with fewer than 100 items identified in the dataset. These elements generally belong to earlier layers of language contact and are less productive in contemporary usage (See Table 2).

<i>agrot'exnik'a</i> 'agrotechnics'	<i>vedra</i> 'bucket'	<i>k'ap'ik'i</i> 'coin'	<i>sk'ovoda</i> 'pan'
<i>ak'ardioni</i> 'accordion'	<i>zondik'i</i> 'umbrella'	<i>k'asruli</i> 'cooking pot'	<i>sp'it'ska</i> 'match'
<i>ak'ovk'a</i> 'window'	<i>zak'ovk'a</i> 'metal rivet'	<i>k'erasink'a</i> 'kerosene lamp'	<i>sp'riezdom</i> 'with meeting'
<i>baradok'i</i> 'mess'	<i>k'uk'la</i> 'doll'	<i>k'let'k'a</i> 'cage'	<i>suxci</i> 'dry'
<i>bulk'i</i> 'bread roll'	<i>k'rask'a</i> 'paint'	<i>p'advali</i> 'cellar'	<i>st'olba</i> 'pillar'
<i>brizent'i</i> 'canvas'	<i>k'amanda</i> 'team'	<i>reik'a</i> 'wooden strip'	<i>st'ovik'a</i> 'stand'
<i>garadok'i</i> 'small town'	<i>k'alts'o</i> 'ring'	<i>sirvst'i</i> 'dampness'	<i>xalt'ura</i> 'side job'
<i>gradi</i> 'hail'	<i>k'alink'a</i> 'raspberry'	<i>set'k'a</i> 'net'	Etc.

Table 2: Russian-Origin Lexical Items (Russizms) in Megrelian

The distribution of russizms is strongly conditioned by sociolinguistic factors, particularly speaker age, topic of conversation and regional background. Their use is more typical of older speakers and individuals originating from Abkhazia (including internally displaced populations), reflecting historical patterns of Russian influence. Overall, the data indicate a clear decline in Russian lexical influence, in contrast to the increasing presence of Georgian-derived and international vocabulary.

4.3 Georgian Lexical Influence

Georgian constitutes the primary source of L2 elements in the corpus. Code-mixing with Georgian is especially frequent in the speech of younger and middle-aged speakers, and its occurrence increases in urban and industrial settlements, as well as in regions with closer contact with Georgian-speaking communities. Economic and professional factors also influence this process (Makharoblidze & Gersamia 2022)

Georgian-derived lexemes in the database include both:

- items without Megrelian equivalents, which fill a lexical gap, e.g. *sagani* ‘item’, *gak’vetili* ‘lesson’, *masts’avlebeli* ‘teacher’ etc., and;
- items with existing Megrelian equivalents, resulting in synonymic parallelism, e.g. *adgili* Megr. equivalent *ak’ani* ‘place’, *axlɔmaxlɔs* Megr. equivalent *xɔɔ-xɔɔs* ‘near-by’, *gadmtsema* Megr. equivalent *gintjama* ‘transmission’ etc..

In the latter case, the choice between Megrelian and Georgian forms is influenced by socio-cultural and pragmatic factors, such as discourse context, speaker identity, and communicative intent.

The data further reveal a distinction between:

- Older, fully integrated borrowings, which have become part of the stable Megrelian lexicon, e.g. *gatjereba* Megrelian equivalent *gatjemeba* ‘stop’, *gzadzvaredini* Megrelian equivalent *zir ak’artu fara* ‘crossroad’, *k’etebe* Megrelian equivalent *ɣɔlama* ‘making’, and;
- Recent, less conventionalized forms, which may still be perceived as foreign or are not yet widely used, e.g. *gviq’vebudes* Megrelian equivalent *metjebudes* ‘someone is telling us’, *gamtknarəns* Megrelian equivalent *gɔkirɔnapuans* ‘makes (someone) yawn’ etc.

This contrast highlights the dynamic nature of lexical borrowing and code-mixing in Megrelian. On the one hand, the continuous influx of Georgian lexical material demonstrates the language’s adaptability; on the other hand, it raises questions about the stability of the native lexical system and may be considered an indicator of language endangerment.

5 Types of Code-Switching and Code-Mixing in Megrelian and their Representation in the Dictionary

The examples extracted from the *MLC* illustrate a range of contact-induced phenomena, including code-mixing, code-switching and lexicalization, which are subsequently processed both lexically and morphologically and incorporated into the Megrelian-English dictionary. For the purposes of this study, three main types are distinguished:

- Insertional mixing, defined as the insertion of L2 elements into an L1 (Megrelian) morphosyntactic frame;

- Alternational switching, referring to the shift between L1 and L2 at the level of clauses or sentence boundaries;
- Lexicalized mixing, involving the sharing and integration of morphosyntactic features across languages, resulting in hybrid forms.

5.1 Types of Code-Mixing

5.1.1 Insertional Mixing

Insertional mixing is the most frequent type observed in the corpus and involves the embedding of Georgian, Russian or international lexical items into Megrelian sentences. Examples 1-3 illustrate different sources of insertion: Georgian, Russian and internationalisms, respectively.

- (1) *baɣanɔbas brei mudgasieni vjapendit, magajtɔ **mgelɔbanas**, daxutʃɔbanas.*
In my childhood, I played many games, for example, game of catch, hide-and-seek.
- (2) *skʷɔɾɔda kuyududa, skʷɔɾɔdat ɔrtəd.*
If she had a gas oven, she baked it on a gas oven.
- (3) *mazia **versia**, namud **intʼernetʼis** gedzu, tis xɔɔ vtɕuank.*
there is another version in the Internet

Example (1) illustrates the insertion of Georgian lexical items into a Megrelian text. Example (2) shows the insertion of Russian-origin items, while Example (3) demonstrates the insertion of a borrowing and an internationalism into a Megrelian text.

The inserted lexical items follow Megrelian morphosyntactic rules, which is consistent with the principles of L2 insertion into a matrix language framework (for the principles, see Myers-Scotton 1993; Auer & Muhamedova 2005). Since the FLEx annotation system involves morpheme segmentation and tagging, the roots of non-Megrelian lexical items in examples 1-3 are marked as L2, while suffixes are analyzed as part of the Megrelian grammatical system. Consequently, dictionary entries are represented in their stem form. Vowel-final nouns are represented with their full phonemic shape in word-final position, as the nominative marker is realized as a zero allomorph. In contrast, the nominative suffix /-i/ of consonant-final nouns is not reflected in the dictionary form (4).

- (4) *mgelɔbana* — L2 game of catch; *daxutʃɔbana* — L2 hide-and-seek; *skʷɔɾɔda* — L2 gas oven; *versia* — L2 version; *intʼernetʼ* — L2 internet etc.

The FLEx database also provides evidence for so-called trigger words (Clyne 2009), i.e. lexical items arising at the intersection of two linguistic systems that may trigger a shift to another language during discourse. These include: (a) proper names, (b) lexical transfers and, (c) bilingual homophones. Among these, bilingual homophones are the least frequent in Megrelian and do not typically develop into borrowings. In some cases, borrowed items coexist with Megrelian equivalents, resulting in synonymic parallelism.

(a) Proper Names. Proper names are not translated, as they retain the same phonological form in both L1 and L2. However, anthroponyms may undergo Megrelian-specific mor-

phological adaptation, which goes beyond purely phonetic processes, e.g. *Nodia – Nodixe*, *Naroushvili – Narouskua*, *Chikava – Chika*, *Lukava – Luka*, etc. (Kartozia et al. 2011). In such cases, they may be treated as instances of code-switching.

(b) Lexical Transfers and Borrowings include items that have no Megrelian equivalent; fulfill a nominative function; are frequently found in technical, administrative, or scientific domains. Borrowings may also occur in cases where a Megrelian equivalent exists, creating a parallel lexical system influenced by sociocultural and pragmatic factors, e.g. *adgili* ‘place’ (129), *adre* ‘early’ (57), *rituali* ‘ritual’ (58), *raioni* ‘district’ (42) etc.

(c) Bilingual Homophones form a distinct group. One group within this category consists of lexical items that function as elements used to organize the process of communication (Matras, 1998). These include discourse markers (words and phrases) such as *anu* ‘that is’, *ese igi* ‘i.e.’, *kho* ‘you know’, *nu* ‘well’, *aba* ‘come on’, *upro* ‘more, mostly’, *ra tkma unda* ‘of course’, *ki batono* ‘certainly’, *magalito* ‘for example’, *vabshe* ‘basically’, *karoche* ‘in short’, and *okey* ‘okay’.

The second group consists of words that have identical orthographic and phonological forms in two different languages but differ in meaning. For example: L1 *axla* ‘now’ vs L2 *ase* ‘thus’; L1 *tu* (conjunction/particle; including the discourse particle /-da/) vs L2 *tu* (conjunction); L1 *goči* ‘piglet’ vs L2 *tu* (conjunction). These items illustrate the complexity of cross-linguistic interaction and the role of phonological overlap in bilingual speech.

5.1.2 Alternational Code-Switching

Alternational code-switching occurs at clause or sentence boundaries, where speakers shift from Megrelian to Georgian (or vice versa) for entire segments of speech (5-6)

- (5) *ipreli gvarťkile, mara p’rakt’ik’uli saubari gvit’irs.*

Megrelian: We understand everything, **Georgian:** *but we have difficulty with practical conversation.*

- (6) *iseti k’argi kali iq’o do tik ma mumonts’q’u sanep’idst’antsias samufad.*

Georgian: She was such a good woman, **Megrelian:** and it was her who helped me find **Georgian:** a job at the **Russian:** sanitary service.

Example (5) represents a complex syntactic construction in which code-switching occurs at the boundary between the first and second clause, shifting from Megrelian to Georgian. The connective *mar* ‘but’, used to link the clauses, belongs to both Megrelian and Georgian dialectal lexical inventories.

In example (6), the direction of switching is reversed: the utterance begins in Georgian (*iseti k’argi kali iq’o* ‘she was such a good woman’), while the subsequent segment continues in Megrelian. Such switching patterns are commonly interpreted in the literature as psycholinguistically motivated (Vogt 1954; Gumperz 1958 and others).

In this case, the use of a formulaic phrase from the dominant language (Georgian) at the beginning of the utterance contributes a particular discursive and expressive effect, add-

ing emphasis and stylistic nuance. It also reflects the speaker's conscious and intentional choice to switch codes, motivated by psycholinguistic factors.

At the end of example (6), a multilingual mixing strategy is observed. The Russian-derived lexical item *sanep'idst'antsia* 'sanitary service' and the Georgian form *samufad* 'for working' represent two distinct linguistic codes within the same utterance. This combination reflects the speaker's self-identification with a multilingual environment.

As it was mentioned above, the final word *samufad* 'for working' is both lexically and morphologically Georgian and appears in the adverbial case, marked by the suffix /-d/. In the FLEx annotation system, such forms are treated as structurally indivisible units, i.e. they are not segmented into morphemes in the analysis field and are represented in the dictionary as complete L2 forms. This principle applies to all units that are lexically and morphologically fully attributable to L2.

5.1.3 Lexicalized mixing

Lexicalized mixing results in the formation of hybrid compounds and morphologically integrated structures that have become conventionalized in Megrelian. In this study, lexicalization is understood as the sharing and integration of morphosyntactic features across languages, affecting both inflectional and derivational bound morphemes. The following discussion focuses on the interaction of such morphemes in the structure of the Megrelian verb.

The Megrelian verb is characterized by a high degree of morphological complexity (Kartozia et al. 2010; Kiria et al., 2021), with most grammatical categories expressed through a series of affixal elements. These bound morphemes are distributed both to the left of the root (prefixal domain) and to the right of the root (suffixal domain).

Prefixal Domain

Negation marker -> Perfective/affirmative marker -> Preverb -> Imperfective marker -> Evidential marker -> Person markers -> Voice / version / causative / potential markers -> R (root)

Suffixal Domain

R (root) -> Augment / thematic marker -> Voice / causative / thematic suffix -> Passive / potential / evidential marker -> Tense-aspect marker -> Paradigm marker / subordinative marker -> Person-number markers -> Emphatic vowel -> Conditional marker -> Clitics (modal particles, conjunctions)

These morphological templates provide the structural framework within which L2 elements are integrated, either through direct insertion or through adapted incorporation. As a result, lexicalized mixing in Megrelian operates not only at the lexical level but also at the morphemic and grammatical levels, demonstrating a high degree of structural integration.

In the FLEx database, the verbal lexicon reveals multiple patterns of interaction between L1 (Megrelian) and L2 (primarily Georgian) morphemes. Two main strategies can be distinguished:

(a) Direct Insertion of L2 Morphemes (6-8)

In this pattern, L2 morphemes (roots, stems or affixes) are inserted into the Megrelian verbal matrix without structural modification. This corresponds to an insertional strategy, whereby the overall morphosyntactic frame remains Megrelian.

(b) Adapted Integration of L2 Morphemes (9-13)

In this case, L2 morphemes are incorporated into the L1 matrix with phonological and morphotactic adaptation, reflecting constraints specific to Megrelian. These adaptations may involve changes in phonological shape or positional behaviour in accordance with Megrelian morphological rules.

The analysis confirms that borrowed elements and mixed forms, even when originating from another language, are systematically shaped by the morphosyntactic structure of the matrix language, in line with the Matrix Language Frame model (Myers-Scotton 1993).

5.1.3.1 Direct Insertion Strategies

The insertional strategy involves the integration of L2 morphemes into the Megrelian (L1) verbal matrix. Examples (7) and (8) demonstrate that, within the prefixal domain of L1 verbs, morphemes may appear that are shared with L2 in terms of phonetic inventory, grammatical function, and structural position (i.e. as prefixes or suffixes). In such cases, the verbal root is selected from the L2 lexicon, while the surrounding morphological structure is largely provided by L1. At the same time, towards the anlaut of the verb, L1 morphemes reappear, resulting in hybrid structures in which the matrix language remains Megrelian, despite the presence of L2 lexical material. In example (7), the root morpheme is of Georgian origin, whereas in example (8) it is derived from Russian:

(7) *a-grdzel-en-k*
 apll-long:L2-ts-s2
 'keep going'

(8) *dɔ-v-t'atfil-an-k*
 prv-s1-sharpen:L2-ts-s1
 'sharpen'

In addition to L2 roots, Megrelian verbal structures may also incorporate L2 preverbs and thematic suffixes. For instance, in example (9), the Georgian preverb-containing stem *ganvitar-* is inserted into the Megrelian verbal matrix, while in example (10), a Georgian root combines with the thematic suffix */-eb/*:

(9) *va-ganvitar-d-u*
 neg-develop:L2-impf-s3sg
 'developed'

(10) *i-kmn-eb-u-u-d-u*
 pass-create-ts:L2-pass-ts:L1-impf-subj3sg
 'to be created'

In example (10), two thematic suffixes co-occur: the Georgian /-eb/ and the Megrelian /-u/. In morphological analysis, both are identified separately. According to the shared grammatical principles of Georgian and Megrelian verbal paradigms, both thematic suffixes are removed in the second series of the TAME system, leaving the verbal roots (e.g., *kmn-*, *q'v-*). This indicates that the Georgian suffix retains an active grammatical function within the paradigm. For this reason, the suffix /-eb/ is included in the dictionary and marked with the tag TS:L2.

In rare cases, L2 morphemes are borrowed together with their grammatical function. Example (11) illustrates the emergence of the Georgian prefix /gv-/ in Megrelian verbal structure. In Georgian, this prefix encodes first person plural object agreement, whereas Megrelian typically employs different grammatical mechanisms to express this category. However, in bilingual speech, this prefix may appear as a functional borrowing:

- (11) *gv-i-q'v-eb-u-u-d-u*
O1pl:L2-appl-naratted:L2-ts:L2-pass-ts-impf-s3sg
 'narrated'

In addition, Georgian preverbs (e.g., *da-*, *a-*, *gan-*, *tʃa-*) may be directly incorporated into Megrelian word structure and are listed in the dictionary as L2 prefixes (preverbs), including their phonetic variants (e.g., *da-* with variants *dɔ-*, *dɛɛ-*; *gan-* with variant *gaan-*).

These forms are derived from examples such as (12) and (13). In example (12), the Georgian preverb /ay-/ is preserved unchanged, corresponding to its use in the Georgian lexeme *ay-sruleba* 'to carry out'. In example (13), the Georgian preverb /ga-/ (as in *ga-vrts'eleba* 'to spread') undergoes phonetic adaptation within Megrelian, following the transformation sequence /ga- -> go- -> gu-/:

- (12) *ay-i-srul-in-u*
prv:L2-pass-carry out:L2-ts-s3sg
 'was carried out'
- (13) *gu-v-a-vrtsel-i-k'ɔ*
prv-s1-apll-spread:L2-pm-cond
 'in order to spread it'

5.1.3.2 Adapted Integration Strategies

Many borrowed or mixed items exhibit partial integration into Megrelian morphology. In such cases, Georgian roots are combined with Megrelian affixes and inflectional endings, resulting in hybrid verbal or nominal forms (14).

- (14) *a-tm-a-b-y-nifn-ɛn-k*
prv-impfv-prv:L2-sub1-prv:L2-celebrate-ts-subj1sg
 "I will mark"

This form combines the Georgian verb root *-nifn-* with Megrelian inflectional morphology. Such examples demonstrate that code-mixing operates at the morphemic level, reflecting systematic grammatical adaptation rather than random alternation. Also, it represents a

more complex type of L1-L2 interaction, in which elements from the embedded language (L2) are not simply inserted into the matrix language (L1) structure, but are restructured in accordance with the morphotactic and phonotactic constraints of Megrelian.

In this example, the Georgian root *-nifn-* appears in its preverb form *ay-nifn-*. According to Megrelian morphotactic rules, the S1 allomorph /b-/ is metathesized into the Georgian preverb /ay-/, effectively splitting it and yielding the form *a-b-y-nifn-*.

The Georgian preverb is thus desemantized, as evidenced by the addition of the Megrelian preverb /a-/ in the anlaut position, which fulfills the same grammatical function. This duplication indicates that the original L2 preverb no longer carries its full functional load within the hybrid structure.

In such hybrid formations, the segments /a/ and /y/ should ideally be treated as components of a single underlying preverb, although this relationship is not fully recoverable in the dictionary representation. Consequently, the lexicographic treatment of such forms cannot always capture the complete morphosyntactic history of the item.

5.2 Combined Code-Mixing

The Megrelian speech of speakers from the city of Poti (Samegrelo region) can be regarded as a particularly illustrative case of combined code-mixing, code-switching and lexicalization, in which the degree of language endangerment becomes especially evident.

Poti is a Black Sea port city that has historically attracted migrant populations from different regions in search of employment. At the same time, it borders the Guria region, where the Gurian dialect of Georgian is spoken. Intensive and prolonged contact between these linguistic communities has fostered frequent code-switching and code-mixing. As a result, Megrelian has largely disappeared from everyday social use in this area and the intergenerational transmission of both the language and its associated cultural practices has been significantly disrupted.

An illustrative example from the Poti variety demonstrates the interaction of multiple types of code-mixing:

- (15) *gamsazyvreli tɛ pakt'is rdu samuɟaɔ adgilebi - p'ɔrt'i, sanavsadguro kalaki, p'irɔbebi.*

'The determining factors of this act were the workplaces - the port, the harbor city, and the working conditions.'

- (16) *p'irɔbeb-i i-kmn-ɛb-u-d-u samuɟaɔ adgilebis.*

'The working conditions would determine the workplaces.'

In example (15), only the demonstrative *tɛ* 'this' and the verb *rdu* 'was' are Megrelian, while the remaining lexical material is of Georgian origin. In contrast, example (16) follows Megrelian syntactic structure, but the verb root and nominal stems are derived from Georgian, with Megrelian morphological markers applied.

Such constructions illustrate how speakers combine Georgian lexical material with Megrelian grammatical templates, resulting in highly hybridized utterances. These combined

patterns of code-mixing, code-switching, and lexicalization reflect not only bilingual competence but also the advanced stage of language shift, in which the structural integrity of the minority language is increasingly reduced.

6 Conclusions

The findings of this study demonstrate that code-mixing in Megrelian is a structured and productive phenomenon, rather than a mere result of lexical borrowing. The analysis of corpus data shows that the integration of L2 elements, primarily from Georgian, is governed by systematic grammatical principles and operates across multiple linguistic levels, including the lexical, morphological and syntactic domains.

The results highlight a high degree of bilingual competence among Megrelian speakers, who are able to combine elements from different linguistic systems while maintaining overall structural coherence. In particular, the data reveal considerable morphological adaptability, whereby Georgian lexical material is incorporated into Megrelian morphosyntactic frameworks. This process is not limited to surface-level insertion, but often involves deeper integration at the morphemic level.

At the same time, the distribution of code-mixing patterns reflects clear sociolinguistic variation. Younger and urban speakers show a higher frequency of mixing, indicating the influence of social, economic, and communicative factors on language use. In certain contexts, especially in environments characterized by intensive language contact, combined forms of code-mixing, code-switching, and lexicalization may also point to an advanced stage of language shift.

From a lexicographic perspective, these findings justify the inclusion of mixed and hybrid forms as dictionary entries. The use of L2 tagging in the Megrelian-English dictionary provides a systematic way to document the origin and structure of such items, thereby offering a more accurate representation of contemporary language use. This approach reflects the linguistic reality of Megrelian, in which multilingual interaction forms an integral part of speakers' everyday communication and identity.

Finally, the data suggest that certain morphemes shared between Megrelian (L1) and Georgian (L2) may be analysed as bilingual homophones, particularly in cases where morphosyntactic markers are phonologically identical, functionally equivalent, and occupy corresponding positions within the morphological structure, for example in prefixal or suffixal chains. This observation further underscores the depth of structural convergence between the two languages.

Overall, the study contributes to the understanding of code-mixing in endangered languages by demonstrating that such phenomena are not merely indicators of language erosion, but also reflect systematic linguistic processes that can be effectively captured through corpus-based and lexicographic methods.

References

- Auer, P. & Muhamedova, R. (2005). Embedded Language and Matrix Language in Insertional Language Mixing: Some Problematic Cases." In *Italian Journal of Linguistics* 17 (1), pp. 35-54.
- Austin, P. K. (2006). Data and Language documentation. In *Essentials of language documentation*, by N. P. Himmelmann, & U. Mosel (Eds.) J. Gippert, pp. 87-113. Berlin & New York: Mouton de Gruyter.
- Bowern, C. (2008). *Linguistic fieldwork: A practical guide*. New York: Palgrave Macmillan.
- Clyne, M. (2009). *Community Languages: The Australian Experience*. Cambridge University Press.
- Gersamia, R. & Lobzhanidze, I. (2021). *Mingrelian Converter*. 03 17. <https://xmf.iliauni.edu.ge/converter/converter.html>.
- Gersamia, R. & Lobzhanidze, I. (2022-2025). *Megrelian Language Corpus (MLC)*. Accessed 02 28, 2026. <https://xmf.iliauni.edu.ge/>.
- Gumperz, J. (1982). *Discourse Strategies*. Cambridge: Cambridge University Press.
- Haugen, E. (1956). *Bilingualism in the Americas: A Bibliography and Research Guide*. Alabama: American Dialect Society.
- Hoo, Y. (2007). Code-mixing: Linguistic Form and Socio-cultural Meaning. In *The International Journal of Language, Culture and Society*, 21, pp. 23-30.
- International, SIL. (2024). *SIL Fieldworks Language Explorer (FLEX)*, Version 9.1. January 01. <https://software.sil.org/fieldworks/>.
- Kartozia, G., Gersamia, R., Lomia, M. & Tskhadaia, T. (2010). *megrulis lingvisturi analizi [Linguistic analysis of Megrelian]*. Tbilisi: Meridiani.
- Kiria, Ch., Ezugbaia, L., Memishishi, O. & Chukhua, M. (2021). *Laz-Megrelian Grammar. I. Morphology [lazur-megruli gramatika. I. morf'ologia]*. Tbilisi: Meridiani.
- Makharoblidze, T. & Gersamia R. (2022). "On Mountainous Samegrelo: General description and linguistic situation." *Vitalité sociolinguistique des langues des massifs montagneux: Alpes et Caucase. 17. Sezione IV. Sociolinguistica*, pp. 303-324.
- Matras, Y. (1998). Utterance Modifiers and Universals of Grammatical Borrowing. In *Linguistics*, 36-2, pp. 281-333.
- Myers-Scotton, C. (1998). *Codes and Consequences: Choosing Linguistic Varieties*. New York: Oxford University Press.
- Myers-Scotton, C. (1993). *Social Motivations for Codeswitching: Evidence from Africa*. Oxford: Clarendon Press.
- Poplack, Sh. (1980). Sometimes I'll Start a Sentence in Spanish y Termino en Espanol: Toward a Typology of Code-switching. In *Linguistics* 18 (233-234), pp. 581-618.

- Poplack, Sh., Sankoff, D. & Miller, Ch. (1988). The Social Correlates and Linguistic Processes of Lexical Borrowing and Assimilation." In *Linguistics* 26 (1), pp. 47-104.
- Redouane, R. (2005). Linguistic Constraints on Codeswitching and Codemixing of Bilingual Moroccan Arabic-French Speakers in Canada. In J. Cohen, K. T. McAlister, K. Rolstad, & J. MacSwan (Eds.), ISB4:. 2005. "Linguistic Constraints on Codeswitching and Codemixing of Bilingual Moroccan Arabic-French Speakers in Canada." *Proceedings of the 4th International Symposium on Bilingualism*. Arizona: Cascadilla Press. 1921-1933.
- Vogt, H. (1954). Language Contacts. In *Word* 10(2-3), pp 365-374.
- Weinreich, U. (1953). *Languages in Contact*. The Hague: Mouton.