Irina Lobzhanidze, Rusudan Gersamia, and Nino Tsulaia

COMPILING A BILINGUAL MEGRELIAN-ENGLISH ONLINE DICTIONARY

Preserving Endangered Kartvelian Languages

Abstract The paper outlines one of the results of the project dedicated to one of the endangered Kartvelian languages, especially Megrelian. Providing data collection and documentation through fieldwork implemented in Samegrelo (Georgia), the project aims to comprehensively document the Megrelian language and encompasses the development of the annotated corpus, sketch grammar, and a bilingual dictionary. As a result, a bilingual Megrelian-English dictionary has been compiled using the Fieldwork Language Explorer (FLeX) and combining technological and traditional lexicographic approaches. We provide numerical examples to highlight the language structure and its application to the compilation of the dictionary, discussing its application to language preservation issues. The paper is subdivided into four parts: 1. Introduction, which outlines the project dedicated to the documentation of Megrelian language within the framework of the project financed by the Shota Rustaveli National Science Foundation (FR-21-993-3, 2021-2025); 2. Lexicographic insights on the Megrelian-English dictionary, which highlights the challenges of preserving endangered Megrelian language; 3. Macro- and micro-structures of the Megrelian-English dictionary, which emphasizes the structure of the dictionary compiled using FLeX and provides information on its licensing and accessibility; 4. Conclusions underscore the importance of this lexicographic effort and its application to the preserving of Megrelian language.

Keywords endangered languages; Kartvelian linguistics; Megrelian language

1. Introduction

The Kartvelian language family is characterized by its relatively uniform sound system, a well-developed system of word inflection and derivation, which involves a large variety of grammatical affixes and internal stem inflection and, the split ergativity of the sentence. Georgian, the most widely spoken Kartvelian language, has a rich literary tradition starting from the fifth century (Chikobava, 2008 [1952]; Shanidze, 1976; Sarjveladze, 1997, and others) and serves as the official language of Georgia. Svan, spoken in the mountainous regions of north-western Georgia, reveals significant phonological and morphological differences from other Kartvelian languages. Megrelian and Laz, spoken in western Georgia and north-eastern Turkey, respectively, share a close relationship and display remarkable similarities in terms of vocabulary and grammar. Despite the significant role played by the Kartvelian languages in shaping the cultural and national identity of the Georgian people, their present situation has undergone notable changes. Especially, Georgian, despite having a relatively small number of speakers, has managed to maintain its resilience. At the same time, other Kartvelian languages such as Svan, Megrelian, and Laz have unfortunately been listed as "increasingly endangered" in the UNESCO Atlas of the

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

World's Languages in Danger (2021). Svan, Megrelian and Laz languages transmit unique cultural knowledge encoding oral tradition that is under the threat of disappearance. The biggest challenge to preserve these unique languages is associated not only with the importance of their documentation, but also with the preparation of a dictionary using the data collected during the fieldworks.

In any case, preserving an endangered Megrelian language presents a lot of challenges provoked by linguistic and sociocultural aspects, including:

- Scarcity of resources: one of the most important challenges is the limited availability of linguistic resources, especially dictionaries. From the contemporary point of view, Megrelian faces a shortage of materials, especially written textual or other media, documented contemporary data and linguistic studies of the above-mentioned data using modern technologies. The absence of a contemporary appropriately annotated Megrelian corpus affects understanding of various linguistic aspects of grammatical structure of Megrelian and its vocabulary. The only one resource available by this moment is a corpus of Megrelian (GNC Megrelian) (Gippert et al., 2011–2024), which includes a small quantity of texts consisting of 89 404 words with punctuation and documented in the previous century without linguistic annotation. All the above-mentioned facts complicates the compilation of grammars, textbooks and dictionaries. As a result, in addition to very old notes on Megrelian like wordlists mentioned in (Gippert, 1992; 2016), the dictionaries of Megrelian are subdivided into printed dictionaries published in the past century (Kipshidze, 1914; Charaia, 1997; Eliava, 1999; Kajaia, 2000–2009, and others) and online dictionaries created during the last time and based on printed dictionaries (Kajaia, 2000–2009; Kobalia, 2010–2020, and others). These dictionaries provide valuable insights about the Megrelian lexicography, but do not capture the contemporary linguistic situation.

- Urgency of the Task: With each passing generation, the number of proficient speakers decreases, making the urgency of the preservation task even more complicated and important. The younger generations do not sufficiently acquire the endangered Megrelian language due to societal and educational influences, leading to a significant gap in generational transmission between the linguistic heritage of older generations and the linguistic proficiency of the younger population and implies that data mentioned above and collected a century ago not only fails to represent the current grammatical structure of a language but also inadequately reflects the present condition of its dictionary.

- Globalisation and Georgian language influence: Increased globalisation and Georgian language influence led to the adoption of the Georgian language and lifestyles, diminishing the value placed on preserving the Megrelian language and leading to a decline in its usage and significance. Addressing this issue requires not only an acknowledgment of the broader sociocultural dynamics influencing language choices but also the implementation of linguistic revitalization efforts, including the creation and compilation of resources (online corpora, grammars, dictionaries, etc.).

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

As a result, a compilation of Megrelian-English dictionary can be considered as an attempt in slowing down the negative impacts on endangered Megrelian making it more accessible worldwide. Its compilation became possible in parallel with the compilation of the annotated online corpus of Megrelian consisting of 97,393 tokens (60,783 types) collected during the language documentation project financed by the Shota Rustaveli National Science Foundation (FR-21-993-3, 2021-2025). The main scope of the project was to collect contemporary data via fieldwork, to process them using the Fieldwork Language Explorer (FLeX) and to compile a corpus, sketch grammar and an online dictionary combining technological and traditional lexicographic approaches. The dictionary is available on the official site of the Corpus of Megrelian language (https://xmf.iliauni.edu.ge/). The resource is intended to serve not only as a comprehensive reference of the contemporary linguistic condition of Megrelian language, but also as a specialized resource for those interested in revitalization of Megrelian language.

2. Language Structure and Lexicographic Insights on the Megrelian-English Dictionary

The compilation of linguistic specification is always closely connected to the macro- and the micro- structures of a dictionary. Deciding on the types of entry the dictionary will include and organizing the headword list are macrostructure decisions, but planning the entries in the dictionary and deciding on their structure and components are microstructure decisions (Atkins & Rundell, 2008).

A single entry can consist of a headword with its accompanying different meanings and, appropriately, it should be stored in the dictionary database (DB) in an alphabetical order. In the case of European normalized languages, the problem of alphabetization cannot be considered as a challenge. In the case of the Megrelian language, the headword list as well as the structure of entries in the dictionary are always under the impact of complex morphological structure, huge amount of word and/or form formation affixes and affect the compilation of the dictionary as a whole.

Thus, the issues under consideration are subdivided into three main parts: nominal and verbal inflections of Megrelian, which cause lemmatization and alphabetization problems, and, correspondingly, the access by end-users to the headwords of dictionary entries.

2.1 Nominal Inflection

In Megrelian, the structures of nouns, adjectives, numerals and pronouns have something in common, but the quantity of slots and formation models are different. Generally, the formation of nominal inflection is carried out by suffixation; only the degrees of an adjective are formed by means of circumfixation.

Types of stems are a base for different types of inflections; therefore, nouns are

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

subdivided into common and proper nouns with different types of inflections. Pronouns are subdivided into personal, demonstrative, indefinite, possessive, interrogative, relative, reciprocal, negative, determinal and reflexive ones; numerals into cardinal, ordinal fractional ones; and adjectives into gradable and non-gradable ones, i.e., those which can produce degrees of comparison and those which cannot.
The main morphological features which affect the formation of nominal inflection are as follows: case (nominative, ergative, dative, genitive, instrumental, adverbial and vocative), number (singular and plural), postpositions associated with concrete cases and clitics. Additional categories are degree for adjectives and person for pronouns. The schemes of nominal formation are as follows:

- Noun: Stem -> Consonant epenthesis -> Number -> Vowel Epenthesis -> Case -> Emphatic Vowel -> Postposition -> Focus -> Emphatic vowel -> Particles[1, 2, 3] -> Conjunction
- Numeral: Stem -> Number -> Vowel Epenthesis -> Case -> Emphatic Vowel -> Postposition -> Focus -> Emphatic vowel -> Particles[1, 2, 3] -> Conjunction
- Pronoun: Stem -> Number -> Vowel Epenthesis -> Case -> Postposition -> Focus -> Emphatic vowel -> Particles[1, 2, 3] -> Conjunction
- Adjective: Degree -> Stem -> Degree -> Consonant epenthesis -> Number -> Vowel epenthesis -> Case -> Emphatic vowel -> Postposition -> Focus -> Particle[1, 2, 3] -> Conjunction.

Using the above-mentioned schemes, it becomes clear that the quantity of slots depends on the part of speech (PoS). Thus, the possible quantity of slots in nominals varies between 11 and 13 slots (1–4).

(1) *t͡ʃxu-l-ɛp-iʃɔ-t-i-a-va-ɔ*
COW-E-PL-BEN-POST-EMPH-QUOT3-QUOT3-QUOT
'As it was said, and for cows'

(2) *t͡ʃit͡ʃɛ-l-ɛp-iʃɔ-t-i-a-va-ni*
little-E-PL-BEN-POST-EMPH-QUOT3-QUOT3-that
'As it was said, and for little ones that'

(3) *vitɔxut-iʃɔ-t-i-a-va-ɔ*
fifteen-E-PL-BEN-POST-EMPH-QUOT3-QUOT3-QUOT
'As it was said, and for fifteen'

(4) a) *atɛna-Ø*
this-SG.NOM
'This'

b) *atɛn-ɛp-iʃɔ-t-i-a-va-ɔ*
this.PROX-PL-BEN-POST-EMPH-QUOT3-QUOT3-QUOT
'As it was said, and for them'

Thus, from the lexicographic point of view, the primary wordform of a dictionary entry is the lemmatized form of a word. In case of nominals, the lemmatized form is a form in the nominative singular, which makes it easy to find nominals by typing the initial letters of a stem. This rule works for adjectives as well, because they are listed in their positive degree.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

2.2 Verbal Inflection

In Megrelian, like in other Kartvelian languages, the main morphological features which affect the formation of verbal inflection are as follows: the TAM (tense-aspect-mood) series, which specify case-marking and linking between participants like agent and patient by means of preverbs, version vowels and thematic suffixes, voice subdivided active, medium and passive, personality, which covers unipersonal, bipersonal and tripersonal verbs, and number. A Megrelian verb contains many morphemes, which from one point of view are typical for agglutinating structures, but from another point of view are characteristic of inflected ones. The Megrelian verbal paradigm can be considered as a mixed one. The scheme of verbal formation is as follows:

Negation particle-> Affirmative particle -> Preverb -> Aspect -> Evidentiality -> Subject&Object agreement -> Applicatives, voice, causation or potentialis -> Stem -> Augment -> Voice, causation -> Thematic suffix or potentialis -> Tense&Aspect -> Subject&Object agreement -> Emphatic vowel -> Mood

Most verbs have preverbs lexically associated with them, although there is a group of verbs that do not have preverbs. Preverbs can be classified as a closed class always associated with verbs (5).

(5)

a) *gɔnɔ-rt-u*	b) *mik'i-rt-u*	c) *ak'ɔ-rt-u*
prv-go-s3sg(aor)	prv-turn-s3sg(aor)	prv-turn-s3sg(aor)
'moved, transferred'	'turned around'	'rushed in, burst in'

In addition, any kind of morpheme gives information about the grammatical function of the word. A Megrelian verb, generally, uses bound morphemes to show its grammatical function. The main types of use are as follows (6–7):

(6) affixation, e.g.,

a) *mɔ-tmɔ-v-ɔ-zim-ap-u-an-d-i-t-i-k'ɔ-ni*
prv-impfv-s1-caus-measure-caus-aug-ts-impf-pm1/2-pl1/2-cond-conj
'if I measured smth.'

b) *gɛmuan-i-ɛ-n-i-a-ni*
delicious-nom-be-stat-s3-emph-quot3-that
'This is delicious'

(7) root alternation, e.g.,

a) *sk'id-u-n-ɔ?*
stay-pass-s3sg-quot(prs)
'Is he/she/it staying?'

b) *i-t'-ɛn-s*
appl.subj-stay-ts-s3sg(prs)
'He/She/It is leaving'

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

The stemming of the verbal form is closely connected to the generation for subject- and object-based paradigms (8–9) and provision of their further analysis.

(8) *v-ɔ-rtsʾq'-ɛ-k*
SUBJ1-APPL.INDIR-see-STAT-SUBJ1SG(PRS)
'sees'

(9) *m-ɔ-rtsʾq'-ɛ-k*
OBJ1-APPL.INDIR-see-STAT-SUBJ1SG(PRS)
'view, observation'

Keeping in mind that the majority of the existing Megrelian dictionaries follow a mixed type of verbal presentation, which covers not only verbal nouns and adjectives (10), but also the third person singular forms (11), both of these forms are represented in the dictionary.

(10) a) *pun-ap-a*
receive-TS-SG.NOM
'receiption'

pun-ap-il-i
receive-TS-PTCP-NOM(PST.PASS)
'received'

b) *ɛ-tʃɔp-u-a*
PRV-take-OBLV-SG.NOM
'taking'

ɛ-ma-tʃɔp-al-i
PRV-PTCP-take-PTCP-NOM(ACT)
'he/she/it who takes smth.'

(11) *ɔ-pun-u-an-s*
APPL.INDIR-boil-AUG-TS-S3SG(PRS)
'boiling'

ɛ-tm-ɛ-tʃɔp-un-s
PRV-PRFV-boil-TS-S3SG(PRS)
'boiled'

From the lexicographic point of view, the problems of lemmatization as well as alphabetization are closely connected to the existence of particles, preverbs, personal markers, applicatives, voice, markers of evidentiality, potentialis and causation, which affect the initial point of a verbal entry and change the lexical meaning of the verb. Personal markers, applicatives and voice markers are generally used to indicate agent, patient and beneficiary and, respectively, subject, direct and indirect object relations, but they do not change the lexical meaning of a verb.

2.3 Lemmatization and Alphabetization

2.3.1 Lemmatizaton

Lemmatization is the process of deriving the base form or lemma of a word from one of its inflected forms and, generally, refers to the language vocabulary. For Kartvelian languages, including Megrelian, lemmatization can be considered a quite difficult task. This difficulty arises because of a huge quantity of agglutinating affixes. For example, a verbal scheme may have at least 8 prefixes associated with separate linguistic features and preceding the root. As a result, the determination of the lemma (i.e., initial point of a word) is a very complicated task.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Additionally, the lemmatization is a part of the linguistic normalization process, which is very complicated in case of endangered low-resourced languages. The linguistic normalization can be thresholds: morphological, syntactical and lexicosemantically. For the lexicographic purposes, the most important is morphological normalization, which strictly follows the Morpho-syntactic Annotation Framework (MAF). Following MAF, the morphological normalization performed by means of lemmatization reduces verbal forms to the infinitive, nominal forms to the nominative singular, comparative and superlative of adjectives to the positive. The only exception described in MAF is a verbal lemma in Arabic represented in the third person singular with the accomplished aspect (ISO/DIS 24611, 2012).

The MAF framework adopted for Modern Georgian can be considered as useful for other Kartvelian languages as well including Megrelian. In case of Megrelian, it means that the lemma for the nominal paradigm is represented with the nominative singular (12), for the adjectives – the prevalence is given to the positive degree (13), while for the verbal paradigm – the lemma can be represented in the form of masdar, while the lexical entry can contain the third person singular in present or future indicative (14) as well.

(12)	a) *k'ɔt͡ʃ-i* man-SG.NOM 'man'	b) *dida-Ø* mother-SG.NOM 'mother'		
(13)	a) *did-i* big-SG.NOM 'big'	b) *mɔ-did-ɛ-Ø* DIM-big-DIM-SG.NOM 'bigger'	c) *ma-did-a-Ø* EQT-big-EQT-SG.NOM 'big like something'	d) *u-did-aʃ-i* SUP-big-SUP-SG.NOM 'the biggest'
(14)	a) *t͡ʃ'k'ɔm-u-a* eat-OBLV-SG.NOM 'eating'	b) *ɔ-t͡ʃ'k'ɔm-u* PRV-eat-TS-S3SG(AOR) 'He/she/it ate'	c) *t͡ʃ'k'ɔm-un-s* eat-TS-S3SG(PRS) 'he/she/it eats'	d) *ɔ-t͡ʃ'k'ɔm-un-s* PRV-eat-TS-S3SG(FUT) 'he/she/it will eat'

All of the above-mentioned approaches affect the compilation of Megrelian-English dictionary.

2.3.2 Alphabetization

The alphabetization problem already described concerning Modern Georgian language (Lobzhanidze, 2018) is relevant for the Megrelian language as well. Following Atkins & Rundell (2008), the problem of alphabetization is relevant to print dictionaries, but not to electronic ones, and it does not exist if all the headwords are single words, but does arise if the headword list contains MWEs. In Megrelian language, like in other Kartvelian languages, the problem arises concerning the verbal forms. In spite of the lemma used in the form of masdar (14), the verbal stem in the verbal template occupies the nineth slot and it makes impossible to arrange verbs in alphabetical order without paying attention to preverbs, prefixal person markers, applicatives etc. As a result, in the case of Megrelian, the alphabetization problem affects not only print but electronic dictionaries as well. In order to allow the user to access the headword of a dictionary entry, the headwords have to be arranged, from one

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

point of view, in accordance with the alphabet and, from another, paying attention to the inflected forms by means of affixation. Keeping in mind that there is a big difference in meaning between verbal forms formed by means of different affixes and the appropriate verbal noun, it is awkward to appropriately decide the alphabetical order of verbs. For example, if we consider the third person singular in the present or future indicative of verbs, the following forms will arise (15):

(15) *mɛ-ur-s* 'goes', *muk'-ur-s* 'passes by', *ɛ-ur-s* 'descends vertically straight', *al-ur-s* 'descends onto something from the side, slantwise', *ɛl-ur-s* 'ascends from the side of something", *ɛʃ-ur-s* 'ascends/rises', *miʃ-ur-s* 'enters the open space in the center', *min-ur-s* 'enters the enclosed space', *gim-ur-s* 'descends' etc.

and the quantity of dictionary entries with affixes attached to the same stem and reflecting different meanings will be more than enough. Otherwise, if we consider the appropriate verbal noun (16):

(16) *-ula* 'going, walking'

to be the headword of the dictionary entry, different meanings of verbal constituents like those mentioned above will not be represented at the appropriate level.

To summarize, the normalization problems like alphabetization and lemmatization in the case of Megrelian are associated with the position of the verbal root in the verbal template.

3. Macro- and Micro-Structures of the Megrelian-English Dictionary

Any kind of electronic dictionary can be described as a database created with purpose to store and provide access to words or multi-word expressions (MWEs) including their meanings and translations. Generally, there are four major prerequisites (Atkins & Rundell, 2008; Gibbon & Van Eynde, 2000, and others) to the design of any dictionary: a) Linguistic specification (of the macrostructure and the microstructure); b) Database management system (DBMS) specification; c) Specification of the phases of lexicographic database construction: input, verification and modification; d) Presentation of and access to lexical information: access, re-formatting, dissemination. In the case of low-resourced endangered languages, three units are connected to the existence of special DBMS, which will allow not only to represent a dictionary but also to document all data including texts aligned with audio/video files and to process them. This system is to be used for the representation of linguistic specification, conversion of data to the appropriate format, its verification and modification and after all, the collected data should be accessed and disseminated online.

3.1 Macro-structure of the Megrelian-English dictionary

Taking into account that Megrelian-English dictionary is a result of a language documentation and corpus creation project, the Fieldworks Language Explorer

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

(FLeX, 2024) was used as DBMS system. FLeX is generally used in the context of lexicography and computational morphology due to its flexibility in configuring corpora, sketch grammars and dictionaries. So, FLeX was considered as a tool to suit the specific needs of the Megrelian language and its structure and to create the initial macro- and micro-structures of the Megrelian-English dictionary. FLeX not only enables the selection of the dictionary type but also allows the customization of the entry structure, including information on items that will be available in the output file for further processing and uploading to the portal.

For the bi-directional Megrelian-English dictionary, we determined the structure of dictionary entries, paying special attention to the output of FLeX. We revised the entries using a corpus-based approach, utilizing the annotated Megrelian corpus compiled in FLeX and consisting of 97,393 tokens (60,783 types). Additionally, we prepared a converter for the FLeX output to make it compatible with the lexical database using Python, converted the data to .sql file, and launched an online version of the dictionary.

3.2 Micro-Structure of the Megrelian-English Dictionary

The structures of dictionaries that could be generated from FLeX are the following: Hybrid forms, Lexeme-based, and Root-based. Given that we compiled the dictionary together with a corpus interface and paying attention to the problems mentioned above, especially lemmatization problems, special attention was paid to lexeme-based and root-based configurations.

- Lexeme-based: In a lexeme-based configuration, complex forms representing a single lexeme or a unit of meaning are used as the main entries. Lexemes are basic units of meaning, including their translations. This configuration is essential for languages with complex word structures, like Megrelian, where a single lexical unit encompasses various phonetic variants of grammatical forms and meanings. Thus, this configuration simplifies navigation by organizing entries based on core meanings (17).

  (17) **iʃɛniʃɛn** (phon. var. **iʃɛɛn; iʃɛiʃɛn**) mod still

- Root-based: In a root-based configuration, root forms with complex forms as subentries are used as the main entries. Roots are considered the main morphemes, and in the case of Megrelian, they can consist of a single vowel or consonant depending on the part of speech. This approach may not be considered user-friendly, but it allows users to search for complex forms stemming from a common root and define the structure of separate words (18). Such an approach was used by Kipshidze in his printed dictionary (Kipshidze, 1914).

  (18) **miʃɑ-** (sp. var. **mʃi-; nʃɑ-;** phon. var. **miʃ-; miʃɛ-; mʃ-; mʃɑ-; məʃ-; məʃɑ-; məʃɛ-; nʃ-; nʃɛ-**) v:Preverb pfx PRV
  miʃartɑp cn turn
  miʃaulɑr ptcp turn

Also, in the case of online dictionary, this approach allows searching not only for complex forms stemming from a common root, but also for separate inflectional morphemes and their meaning (19).

(19) **-iʃɔ** (dial. var. **-iiʃɔ**, phon. var. **-əʃɔ**) 1) n: Case sfx BEN 2) num: Case sfx BEN 3) adj: Case sfx BEN 4) pro-form: Case sfx BEN 5) ptcp: Case sfx BEN

The FLeX exporting function allows different options, especially, data can be represented as a) Configured Dictionary – Web page (XHTML); b) Dictionary, Reversal index – Webonary; c) Filtered Lexicon – LIFT 0.13 XML; d) Full Lexicon – LIFT 0.13 XML; e) Full Lexicon (lexeme-based) – SFM and f) Full Lexicon (root-based) – SFM. Taking into account that the lexeme- and root-based configurations were chosen for the online Megrelian-English dictionary, the special attention was paid to two formats:

1) Full Lexicon (lexeme-based) – SFM format, which allows exporting of the dictionary using Dictionary Formatter (MDF) lexeme-based standard (20):

(20)
\lx xʃirɑs
\lx_xmf xʃirɑs
\sn 1
\ps_en Temporal
\ps_kat drɔis zmnizɛdɑ
\g_en often
\sn 2
\ps_en Temporal
\ps_kat drɔis zmnizɛdɑ
\g_en frequently

2) Full Lexicon (root-based) – SFM, which provides exporting of the full lexicon using the Multi-Dictionary Formatter (MDF) root-based standard. In this format, subentries are included as part of the main entry rather than as separate entries with links to them (21).

(21)
\lx nɑt͡svl
\lx_xmf nɑt͡svl
\sn 1
\ps_en Main verb
\ps_kat mtɑvɑri zmnɑ
\g_en replace

Both formats are compatible with Lexique Pro for publishing dictionaries online or in print. Additionally, both formats can be easily transformed into .sql format, which is important for integrating the dictionary into the portal. The transformation was made by a Python script specially developed for these purposes. A dictionary database (.db) file is linked to a FLeX corpus and its online version also is connected to the annotated Megrelian corpus interface. The dictionary database includes several key units of information, including the following core units:

- Lexeme or Root Form: The lexeme or root form represents the basic, uninflected or unmodified form of a word. It is the root or the lexeme form that serves as a reference point for various inflections, derivations, or variations of the word.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

This form is, also, used for alphabetization purposes and allows users to look for the entries by pressing alphabet letters as well as to use the search option to look through the entries in alphabetical order.

- IPA (International Phonetic Alphabet): The IPA unit includes the phonetic transcription of the lexeme, allowing users to accurately pronounce the word.
- Gloss: This unit provides a brief, user-friendly explanation or translation of the meaning of a lexeme or a root into English. It serves as a quick reference in English for users to understand the sense of the word.
- Grammatical Information (Part of Speech): This unit specifies the grammatical category or part of speech to which the lexeme or the root belongs. It helps the user to understand the word's syntactic function and to look through its grammatical behaviour.
- Sense: The sense unit provides a detailed explanation of the different meanings or senses associated with a lexeme.

The connection between the dictionary and the corpus enriches the lexical entries with word usage context. The corpus interface allows users to explore how words are used in contextual variations across different genres and registers and, to determine their usage frequency. By analysing the occurrences of a word within the corpus, the users identify common and less common usages, especially in case of code-switches. This information helps researchers to identify the existence and importance of a word in everyday life. To summarize, at this moment the dictionary is available online at https://xmf.iliauni.edu.ge/ and the connection between the dictionary and the corpus interface contributes to deep analysis of language patterns allowing users not only to see the meaning of words but also their morphosyntactic features.

3.3 Potential Users and Language Preservation Issues

Taking into account that the main scope of the project was to collect contemporary Megrelian data and to create an annotated corpus of Megrelian language with an online Megrelian-English dictionary, the potential users of the whole resource are diverse and include Megrelian speakers, who want to preserve their language and use vocabulary, language learners, who need a reliable resource for learning and translating Megrelian, linguists, researchers and students, who learn the grammatical structure of Megrelian and its contemporary situation. Also, the preservation issues can be ensured by regular updates to the corpus and the dictionary, compiling the comprehensive entries and maintaining traditional linguistic elements.

4. Conclusions

In conclusion, this lexicographic effort directed to the maintenance of low-resourced endangered Megrelian language is crucial for its preservation. Through the documentation of Megrelian language, lexicography becomes a keeper of linguistic heritage, collecting, categorizing, and preserving valuable linguistic data to prevent

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

its loss. The urgency of preserving Megrelian is provoked by the declining number of proficient speakers with each passing generation.

Also, the development of Megrelian-English dictionary plays an important role not only in the revitalization of Megrelian but also in the globalisation of project results. This dictionary freely accessible online can be considered as an important learning resource for those interested in its learning not only on the territory of Georgia, but worldwide.

The linking of the dictionary to the annotated corpus of Megrelian allows users not only to look through the dictionary entries but also to read different contexts expressing cultural knowledge, traditions, and identity. Such approach contributes to the cultural preservation of Megrelian as well. While from a linguistic perspective, this recourse contributes to the learning of language structures, grammar, and vocabulary.

References

Atkins, S., & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography.* Oxford University Press.

Charaia, P. (1997). *Megrelian-Georgian Dictionary.* SPB.

Chikobava, A. (2008 [1952]). *enat'mec'nierebis šesavali (Introduction to linguistics).* t'bilisis saxelmcip'o universiteti (Tbilisi State University).

Eliava, G. (1999). *Megrelian-Georgian Dictionary.* Intellect.

Gibbon, D., & Van Eynde, F. (2000). *Lexicon Development for Speech and Language Processing.* Kluwer Academic Publishers.

Gippert, J. (1992). The Caucasian language material in Evliya Çelebi's 'Travel Book'. A Revision. In G. Hewitt (Ed.), *Caucasian Perspectives* (pp. 8–62). Lincom.

Gippert, J. (2016). Complex Morphology and Its Impact on Lexicology: The Kartvelian Case. In T. Margalitadze, & G. Meladze (Eds.), *The XVII EURALEX International Congress* (pp. 16–37). Ivane Javakhishvili Tbilisi University Press.

Gippert, J., Meurer, P., & Tandashvili, M. (2011–2024). *The Georgian national corpus.* Retrieved from The Georgian National Corpus: http://gnc.gov.ge/

Glonti, A. (1964). *k'art'uli enis lek'sikologia (Georgian lexicology).* c'odna (Knowledge).

Glonti, A. (1983). *k'art'uli lek'sikograp'iis sakit'xebi (Questions in Georgian lexicography).* sabčot'a sak'art'velo (Soviet Georgia).

Kajaia, O. (2000–2009). *Megrelian-Georgian Dictionary, 4 volumes.* Nekeri.

Kipshidze, I. (1914). *Megrelian-Russian Dictionary.* Typography of the Imperial Academy of Sciences.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Klimov, G., & Kajaia, O. (2023). *Megrelian-Russian-Georgian Dictionary.* Govorun.

Kobalia, A. (2010–2020). *Megrelian Dictionary.* Artanuji.

Lobzhanidze, I. (2019). Computational Model of the Modern Georgian Language and Search Patterns for an Online Dictionary of Idioms. *Language, Logic, and Computation. TbiLLC 2018. Lecture Notes in Computer Science, vol 11456*, 187–208.

Pochkhua, B. (1974). *k'art'uli enis lek'sikologia (Georgian Lexicology).* t'bilisis saxelmcip'o universiteti (Tbilisi State University).

Sarjveladze, Z. (1997). *žveli k'art'uli ena (Old Georgian Language).* Tbilisi State Pedagogical University Press.

Shanidze, A. (1976). *žveli k'art'uli ena (Old Georgian Language).* t'bilisis saxelmcip'o universiteti (Tbilisi State University).

SIL Feildworks Language Explorer (FLeX), Version 9.1. Retrieved January 1, 2024, from https://software.sil.org/fieldworks/

Standard ISO/DIS 24611 (2012). *Language Resource Management – Morpho-syntactic Annotation Framework (MAF), No 24611.* https://www.iso.org/standard/51934.html

UNESCO: Moseley, Ch. & Nicolas, A. (2021). *World Atlas of Languages: Mingrelian.* Retrieved from https://en.wal.unesco.org/languages/mingrelian

Acknowledgements

This work was supported by Shota Rustaveli National Science Foundation of Georgia (SRNSFG) [grant number: FR-21-993-3, 2021-2025]. The authors would like to express their gratitude to the two anonymous reviewers for the many helpful comments and suggestions and to the organizing committee of the EURALEX for their support.

Contact information

Irina Lobzhanidze
Ilia State University, Institute of Linguistic Studies
irina_lobzhanidze@iliauni.edu.ge

Rusudan Gersamia
Ilia State University, Institute of Linguistic Studies
rgersamia@iliauni.edu.ge

Nino Tsulaia
Ilia State University, Institute of Linguistic Studies
nino.tsulaia.1@iliauni.edu.ge