Preserving Endangered Kartvelian Languages: Lexicographic Insights for the Megrelian Dictionary

Irina Lobzhanidze, Rusudan Gersamia
Ilia State University
E-mail: irina_lobzhanidze@iliauni.edu.ge, rgersamia@iliauni.edu.ge

Abstract & Acknowledgements

The preservation of endangered languages is an important task in the field of linguistic and cultural heritage preservation. This paper focuses on preserving the Megrelian language, an endangered Kartvelian language, through a compilation of the Megrelian-English dictionary using the Fieldwork Language Explorer (FLeX). The study aims to develop a bilingual Megrelian-English dictionary and underscores the significance of this lexicographic effort in both documenting and potentially preserving this endangered language. By integrating modern technology with traditional lexicographic methods, the study offers valuable insights into the preservation and study of endangered Kartvelian languages. The paper is subdivided into the following parts: 1. Introduction; 2. Lexicographic insights on developing dictionaries for the endangered Kartvelian languages; 3. Macro – and microstructures of the lexicographic database; 4. Findings, and conclusions. In summary, the paper demonstrates that electronic lexicography can play an important role in preserving endangered Kartvelian languages.

This paper presents the partial results of a project dedicated to compiling the Annotated Megrelian Corpus with formal grammar and an electronic dictionary, supported by the Shota Rustaveli National Science Foundation (Nos. FR-21-993).

Keywords: Megrelian language; language preservation; electronic lexicography.

1. Introduction

By creating dictionaries for endangered languages like Megrelian, linguists and the language community establish a vital resource that serves the following purposes:

Documentation: Lexicography enables the collection, categorization, and preservation of linguistic data, ensuring that these languages are not lost.

Revitalization: Lexicographic efforts, such as the development of Megrelian-English dictionaries, provide a basis for language revitalization. When these resources are made accessible to the speakers and learners of the language, they become valuable tools for education and language revival.

Cultural Preservation: Languages are repositories of cultural knowledge, traditions, and identity. Lexicography allows for the preservation of cultural heritage, ensuring that

the unique expressions, stories, and worldviews embedded in these languages are passed down to future generations.

Linguistic Research: Lexicographic work on endangered languages contributes to the broader field of linguistics. It aids in the understanding of language structures, grammar, and vocabulary, shedding light on the linguistic diversity of the world.

Online Access: The creation of online dictionaries for endangered languages brings global access to the data, promotes linguistic diversity and supports language preservation.

Any kind of electronic dictionary can be considered a database; generally, its purpose is to provide adequate explanation or translation of separate words or multi-word expressions (MWEs), to store information and to allow the user to find appropriate language units. Following Atkins & Rundell (Atkins & Rundell, 2008), Gibbon & Van Eynde (Gibbon & Van Eynde, 2000) and others, there are four major prerequisites to the design of any lexicographic database, i.e., a dictionary:

- Linguistic specification (of the macrostructure and the microstructure);
- Database management system (DBMS) specification;
- Specification of the phases of lexicographic database construction: input, verification and modification;
- Presentation of and access to lexical information: access, re-formatting, dissemination.

In the case of the Megrelian language, the compilation of a Megrelian-English dictionary using the Fieldwork Language Explorer (FLeX) exemplifies the combination of modern technology with traditional lexicographic approaches. The linguistic specification was developed following the linguistic peculiarities of the Megrelian language; the FLeX allows the exporting of the full lexicon according to the configuration chosen in a format that can be easily transformed to the input required for the online dictionary; all phases of the lexicographic database construction were implemented according to the needs of the project and at this stage, the system undergoes appropriate verification and modifications. This approach not only facilitates the preservation of Megrelian but also offers a template for other endangered Kartvelian languages to follow. The use of technology enhances the accessibility and usability of this dictionary, ensuring that it can reach a wider audience.

2. Lexicographic insights on developing dictionaries for the endangered Kartvelian languages

Preserving an endangered language, particularly one from the Kartvelian family, presents a lot of challenges. These challenges stem from both the linguistic and sociocultural aspects of these languages:

- Scarcity of Resources: One of the foremost challenges is the limited availability of linguistic resources. This scarcity of existing materials, online dictionaries, and contemporary linguistic documentation of the modern situation complicates the preservation efforts. From the contemporary point of view the Megrelian language faces a shortage of existing materials, especially, written documentation and linguistic studies using modern technologies.

The absence of a contemporary Megrelian corpus affects linguistic research and the understanding of various linguistic aspects not only of grammatical structure but also of its vocabulary.

The existing dictionaries of Megrelian can be subdivided into printed dictionaries published in the past century (Kipshidze, 1914, Charaia, 1997, Eliava 1999, and others) and online dictionaries created in the previous decade and based on printed dictionaries (Kajaia, 2000-2009, Kobalia, 2010-2013; Giorgashvili, 2022 and others). These dictionaries provide valuable insights into the development of 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 important. Also, 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 collected a century ago not only fails to represent the current grammatical structure of the language but also inadequately reflects the present condition of its dictionary. This challenge presupposes the need to document the language before it faces deeper changes.
- 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 corpus, dictionary etc.).

Addressing these challenges requires different approaches and lexicography plays an important role in language revitalization. It is well-known that one of the crucial points of language revitalization involves the creation and compilation of appropriate educational resources like grammar, dictionaries, etc. Such resources mix modern technologies and traditional approaches to linguistic data and include the creation and development of online corpora and dictionaries. In the case of Megrelian, the urgent task was to develop an annotated corpus of Megrelian and different configurations of online dictionaries linked to the above-mentioned annotated corpus.

3. Macro- and Micro-structures of the lexicographic database

And as a result the compilation of Megrelian-English dictionary became possible in parallel with the compilation of the annotated online corpus of Megrelian consisting of 97479 tokens (60661 types) collected during the language documentation project financed by the Shota Rustaveli National Science Foundation (project No FR-21-993-3, 2021-2025). The main scope of the project was to collect contemporary data on Megrelian via fieldwork. Ex-

peditions have been carried out every summer since the beginning of the project, focusing on the audio recording of language speakers from different villages within the Samegrelo district. The collected audio files have been converted to text and processed using FLeX and then ELAN, and these have been used to compile a corpus, sketch grammar, and an online dictionary, combining technological and traditional lexicographic approaches. The corpus mark-up strictly follows Leipzig Glossing Rules (Comrie et al. 2008) and Eurotyp Guidelines (Bakker et al. 1993) and includes information on Part of Speeches (PoS) as well as their morphological features.

And, the characteristics of the dictionary comprising 7840 entries can be considered as a part of a database used to store information and provide access to words or multi-word expressions. So, the design of the macro-/microstructure of a lexicographic database for the Megrelian language, i.e., an electronic dictionary, depends on its linguistic features. Considering that the compilation of the Megrelian-English dictionary is a result of a language documentation and corpus creation project, the primary objective was to grant end-users access to the headwords of dictionary entries using the output of the linguistic software FLeX. FLeX is generally used in the context of lexicography and computational morphology due to its flexibility in configuring dictionaries to suit the specific needs of the Megrelian language and its structure. FLeX not only enables the selection of the dictionary type but also allows the customization of the entry structure (see, Fig. 1), including information on items that will be available in the output file for further processing and uploading to the web-site.

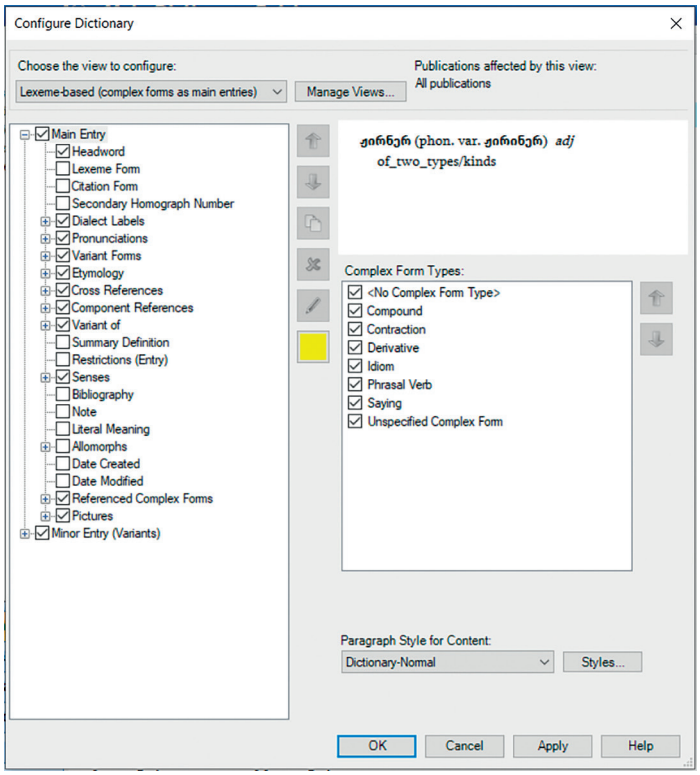

Figure 1: *Determining entry structure*

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. Additionally, we prepared a converter for the FLeX output to make it compatible with the lexical database using Python, created an .sql file, and launched an online version of the dictionary.

The types 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, we chose two types of configurations more applicable in the case of Megrelian, especially Lexeme-based and Root-based. Both of these types will be available online after the implementation of the project:

Lexeme-based (see, Fig. 2): 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.

იშეენ phon. var. of **იშენიშენ**
იშეიშენ sp. var. of **იშეიშენი**, phon. var. of **იშენიშენ**
იშეიშენი (sp. var. **იშეიშენ**, unspec. var. **იშეჲშენ**, phon. var. **იშეჲშენი**) *mod* not_that_much
იშენ sp. var. of **იშენი**
იშენ იშენ
იშენი (sp. var. **იშენ**; **იშენჳ**) *foc* **1)** anyway **2)** at_least
იშენიშენ (phon. var. **იშეენ**; **იშეიშენ**) *mod* still
იშენჳ sp. var. of **იშენი**
იშეჲშენ unspec. var. of **იშეიშენი**
იშეჲშენი phon. var. of **იშეიშენი**

Figure 2: *Lexeme-based configuration*

Root-based (see Fig. 3): 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. Such an approach was used by Kipshidze in his printed dictionary published in 1914.

მიმა- (sp. var. **მმი-**; **ნმა-**, phon. var. **მიმ-**; **მიმე-**; **მმ-**; **მმა-**; **მჳმა-**; **ნმ-**; **ნმე-**) *v: Preverb pfx* PRV
მიმართაფ *cn* turn
მიმაუღარ *cn* turn
მიმაყონაფა *ger* bringing_sb/sth_in
მიმე- (dial. var. **მიიმე-**, phon. var. **მუმე-**) phon. var. of **მიმა-**

Figure 3: *Root-based configuration*

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.

ი-$_2$ (phon. var. **^0-**$_1$; **ჲ-**$_2$; **ჳ-**) *v: Voice, Causation pfx* PASS
ი-$_3$ (dial. var. **იი-**, sp. var. **^0-**$_1$, phon. var. **ჲ-**$_1$; **ჳ-**) *v: Applicatives, Voice, Causation, Potentialis pfx* APPL.OBJ1/2
ი-$_1$ (dial. var. **იი-**, sp. var. **ჳ-**, phon. var. **^0-**$_2$; **ე-**$_3$; **ჲ-**$_1$) phon. var. of **ჱ-**$_1$ (phon. var. of **ქო-**)
ი-$_4$ (dial. var. **-ი**$_1$) *v pfx* E

Figure 4: *Inflectional morphemes*

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) – Standard Format Multi-Dictionary (SFM) and f) Full Lexicon (root-based) – SFM (See, Fig. 4).

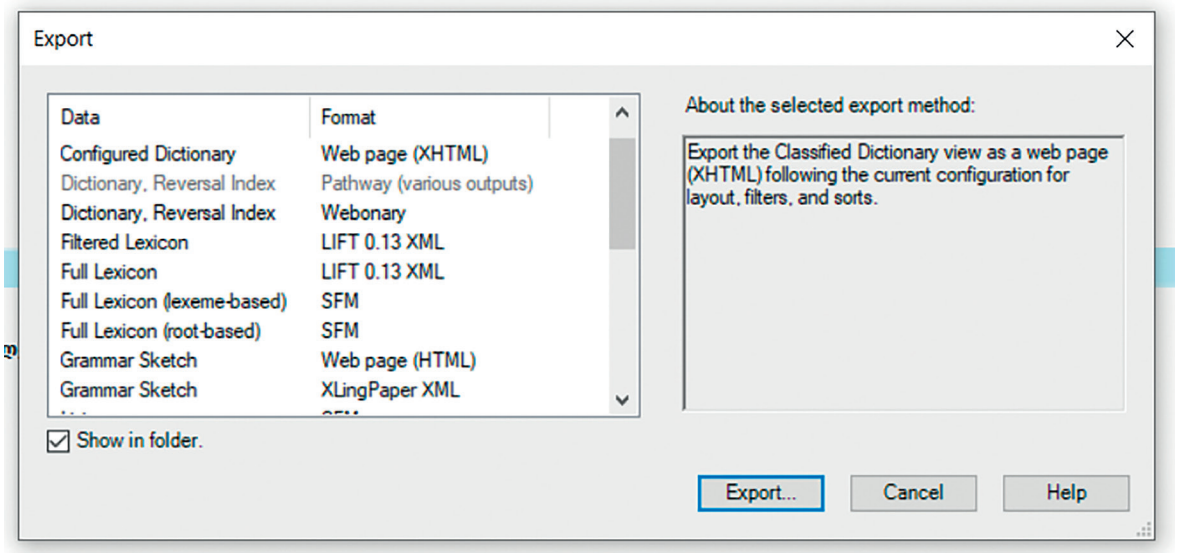

Figure 5: *Dictionary export options*

For the Online Megrelian-English dictionary, Lexeme-based and Root-based configurations were chosen. Special attention was given to the following .db outputs:

Full Lexicon (lexeme-based) – SFM, which exports the full lexicon using the Multi-Dictionary Formatter (MDF) lexeme-based standard. Subentries are exported as separate entries.

(1) \lx ხშირას
\lx_xmf ხშირას
\sn 1
\ps_en Temporal
\ps_kat დროის ზმნიზედა
\g_en often
\sn 2
\ps_en Temporal
\ps_kat დროის ზმნიზედა
\g_en frequently

Full Lexicon (root-based) – SFM, which exports 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.

(2) \lx ნაცვლ
\lx_xmf ნაცვლ
\sn 1
\ps_en Main verb
\ps_kat მთავარი ზმნა
\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 (see, Fig. 5), which is important for integrating the dictionary into the portal. The transformation was made by a Python script specially developed for these purposes.

("ხშირას", "xʃirɑs", "1", "Temporal", "often"),
("ხშირას", "xʃirɑs", "2", "Temporal", "frequently"),
("დრო", "drɔ", "1", "Common Noun", "time"),

Figure 6: *.sql format after transformation*

A dictionary database (.db) file is linked to a FLeX corpus and its online version is also connected to the annotated Megrelian corpus interface. The dictionary database (see, Fig. 6) 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 lexeme form that serves as a reference point for

various inflections, derivations, or variations of the word. 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 lemma.

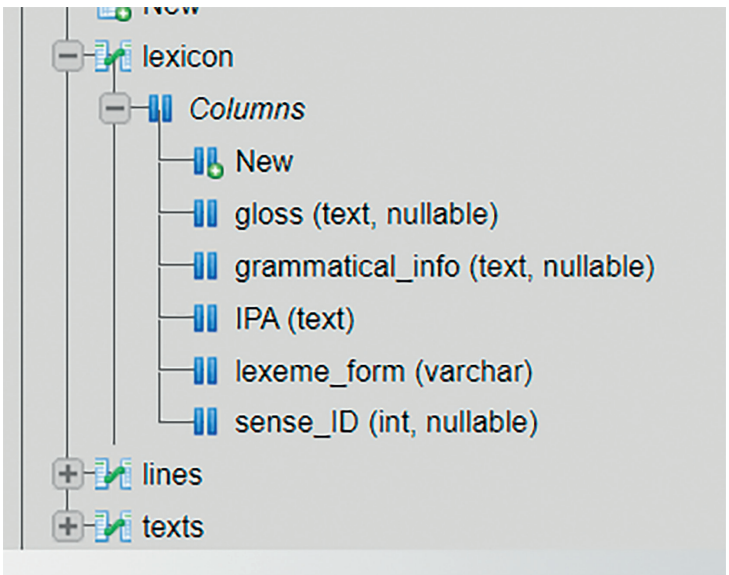

Figure 7: *Database entries*

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 analyzing 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, 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. And, as a result potential users of Megrelian-English

dictionary include, but are not limited to Megrelian speakers aiming to preserve their language, language learners seeking reliable resources, and linguists, researchers, and students studying Megrelian grammar and its current sociolinguistic situation.

4. Findings and conclusions

In conclusion, the lexicographic approaches to low-resourced languages like Megrelian can be considered as an important effort for language preservation. Firstly, through language documentation, lexicography becomes a keeper of linguistic heritage, collecting, categorizing, and preserving valuable linguistic data to prevent the loss of low-resourced endangered languages. The urgency of this preservation task concerning Megrelian is underscored by the declining number of proficient speakers with each passing generation.

Moreover, the development of the Megrelian-English dictionary plays a central role not only in the revitalization of Megrelian, but also in the globalisation of project results. This dictionary can serve as a fundamental tool for language learning purposes when made freely accessible online to speakers and learners.

The fact that the dictionary is linked to the annotated corpus of Megrelian allows users to look through not only the dictionary entries, but also to read real-world different contexts expressing cultural knowledge, traditions, and identity. By preserving these data, the compilation of Megrelian-English dictionary contributes to the cultural preservation of Megrelian.

From a linguistic research perspective, the compilation of a Megrelian-English dictionary linked to a corpus contributes valuable insights into language structures, grammar, and vocabulary.

The use of technology during the compilation of Megrelian-English dictionary, especially, FLeX, made it possible to combine traditional and modern lexicographic approaches and allowed the creation of an online resource for one of the low-resourced Kartvelian languages. The comprehensive lexicographic database linked to a corpus interface serves as a dynamic resource that not only aids language learners and researchers but also contributes significantly to the preservation and revitalization of Megrelian and pays special attention to the use of separate words in context and the linguistic features of concrete morphemes. The compilation of the Megrelian-English dictionary is one of the project results, which contributes significantly to both the preservation of the language and the broader understanding of the Megrelian culture. The data collected has not only provided a comprehensive record of the vocabulary but has also illuminated the grammatical structures and semantic richness of Megrelian.

5. References

Association, I. P. (1999). *Handbook of the International Phonetic Association: a guide to use of the International Phonetic Alphabet.* Cambridge: Cambridge University Press. Retrieved July 16, 2019, from https://www.internationalphoneticassociation.org/

Atkins, S. B. T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography.* NY: Oxford University Press.

Bakker, Dik, Dahl, Östen, Haspelmath, Martin, Koptjevskaja-Tamm, Maria, Lehmann, Christian, and Siewierska, Anna. (1993). *Eurotyp Guidelines.* European Science Foundation in Language Typology.

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

Chikobava, A. (1938). *Chanur-Megrelian-Georgian Comparative Dictionary.* Tbilisi: Academy of Sciences.

Comrie, Bernard, Haspelmath, Martin, and Bickel, Balthasar. (2008). *The Leipzig Glossing Rules: Conventions for Interlinear Morpheme-by-morpheme Glosses.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dilman, D. (1978). *Mail and Telephone Surveys: The Total Design Method.* New Jersey: Wiley.

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

Gersamia, Rusudan, Lobzhanidze, Irina. (2023, 03 17). *Mingrelian Converter*. Retrieved from Irina Lobzhanidze's Personal Website: https://irinalobzhanidze.com/megrelian/converter/converter.html

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

Kadzhaia, O., Fähnrich, H. (2001). *Mingrelisch-Deutsches Wörterbuch.* Reihe: Kaukasienstudien Band 03.

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

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

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

Kobalia, A. (2010-2020). *Megrelian Dictionary.* Tbilisi: Artanuji.

Kurdadze, R., Shonia, D., Tandilava, L., Nizharadze, L. (2015). *Georgian-Megrelian-Laz-Svan-English Dictionary.* Tbilisi: Petite.

ELAN, Version 6.8 (2024, January 01). Retrieved from The Language Archive: https://archive.mpi.nl/tla/elan/download

SIL Fieldworks Language Explorer (FLeX), Version 9.1. (2024, January 01). Retrieved from Fieldwork: Language Technology: https://software.sil.org/fieldworks/

Standardization, I. O. (1996). *Information and documentation — Transliteration of Georgian characters into Latin characters.* https://www.iso.org/standard/17892.html.