The Megrelian Language Corpus (MLC): Creation, Annotation, and Initial Steps Toward a UD Treebank

Irina Lobzhanidze[1], **Rusudan Gersamia**[2], **Tamar Gogia**[3]
Ilia State University, Georgia[1], Ilia State University, Georgia[2], Pompeu Fabra University, Spain[2]
irina_lobzhanidze@iliauni.edu.ge[1], rgersamia@iliauni.edu.ge[2], author3@hhh.com[3]

Abstract

This paper presents the development of the Megrelian Language Corpus (MLC), a new language resource for the documentation and computational analysis of Megrelian, an endangered Kartvelian language. The corpus is based on fieldwork conducted in Samegrelo, Georgia (2022–2024) and currently contains 97,691 tokens and 60,959 types. The data were transcribed using the International Phonetic Alphabet (IPA) and annotated in Fieldworks Language Explorer (FLEx) with segmentation, morphological analysis and bilingual Georgian-English translations. Each text is accessible through a specially designed web interface, providing multiple tiers of annotation and integrated search functions. The paper describes the corpus design, annotation methodology and challenges encountered in representing Megrelian's complex agglutinative morphology. It also outlines initial steps toward converting existing data into the Universal Dependencies (UD) framework, building on experience from related Kartvelian languages such as Georgian. The MLC corpus represents the first publicly available linguistic resource for Megrelian and provides a foundation for future UD treebank development.

Keywords: Megrelian, endangered languages, Universal Dependencies

1. Introduction

The spoken and written corpora are essential resources for the documentation and analysis of any language. However, the development of such corpora for endangered and low-resourced languages remains a major challenge, often constrained by the scarcity of standardized orthography, trained annotators and technological support. The Megrelian language (ISO 639-3: xmf), classified as an "increasingly endangered" member of the Kartvelian family in the UNESCO Atlas of the World's Languages in Danger (UNESCO, 2021), illustrates the problems just described. Spoken primarily in western Georgia, Megrelian does not have an official written status and is used in oral communication (Kartozia et al., 2010). Its complex agglutinative and inflectional morphology further complicates automatic analysis and annotation (Gersamia, 2020).

To address this gap, the Megrelian Language Corpus (MLC) by Gersamia and Lobzhanidze (2022–2025b) has been developed as part of a larger documentation initiative funded by the Shota Rustaveli National Science Foundation (project No FR-21-993-3, 2021-2025). The corpus includes both spoken and written data, collected through contemporary fieldwork (2022-2024) in the Samegrelo-Zemo Svaneti region. The audio recordings were transcribed into the International Phonetic Alphabet (IPA). Morphosyntactic annotation and Megrelian-Georgian-English translations were carried out in FLEx (SIL International, 2025). Video recordings, as well as FLEx-annotated files, were further processed and time-aligned in ELAN (Max Planck Institute for Psycholinguistics, 2025). The resulting resource comprises 97,691 tokens and 60,959 types, forming the first publicly available linguistic data set for Megrelian. The corpus is currently accessible through the web interface and is licensed under the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International (CC BY-NC-SA 4.0) license. The ELAN annotation files are available for download. Future releases will also include downloadable versions of the linguistically annotated text files in .xml format, making the corpus suitable for large-scale computational and machine learning applications.

This paper describes the design, data collection procedures and annotation of the MLC corpus available at `https://xmf.iliauni.edu.ge/` and describes initial steps toward mapping its existing annotation to the Universal Dependencies (Nivre et al., 2017) framework. Section 2 presents the corpus within existing work on Megrelian and other Kartvelian language resources, highlighting the absence of contemporary, linguistically annotated datasets. Section 3 describes the compilation workflow, including fieldwork, transcription practices, morphological annotation in FLEx and database. Section 4 introduces the rationale and procedure for mapping the existing annotation to the UD framework, describing challenges involved in adapting UD to agglutinative morphology of Megrelian. In addition, Section 5 discusses current limitations and future directions, including the expansion toward a fully developed UD treebank.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3250–3256
11–16 May 2026.

2. Related Work

Research on Megrelian has a long descriptive and lexicographic tradition. Early documentation by (Lamberti, 1654; Güldenstädt, 1787–1791; Klaproth, 2012) provided the first grammatical observations and lexical records of the language, forming the historical foundation for language studies. Later linguistic investigations by (Tsagareli, 1880; Chikobava, 1936; Machavariani, 2002; Danelia and Dundua, 2006; Kartozia, 2008) provided detailed descriptions of phonology, morphology and syntax. In spite of the fact that these descriptive works remain important for theoretical and comparative studies, they were not designed to be considered as basis for machine-readable corpora.

Lexicographic efforts have also contributed substantially to the documentation of the language. The printed dictionaries by (Kipshidze, 1914; Eliava, 1997; Kilanava, 2010) and later online editions by (Kajaia, 2000–2009; Kobalia, 2010–2020) provided bilingual mapping and expanded sense inventory. However, these resources lack aligned audio recordings, morphosyntactic annotations and are limited concerning their application for corpus-based linguistic analysis.

Digital corpus initiatives have partially addressed the limitations mentioned above, especially, the Georgian National Corpus (GNC) (Gippert et al., 2011–2025) and the ARMAZI (Gippert et al., 2016) database contain only a small subset of Megrelian and Svan materials, lacking morphosyntactic annotation, time-aligned audio/video recordings and contemporary spoken material. As a result, they do not support computational modeling or the development of dependency treebanks for Megrelian.

The present study directly addresses these gaps by developing the MLC, the first linguistically annotated corpus for Megrelian that integrates both spoken and written data collected through recent fieldwork. Unlike earlier descriptive or lexicographic resources, the MLC provides morphosyntactic annotation, aligned bilingual translations and structured XML data suitable for computational processing. Furthermore, while Universal Dependencies (UD) representation has only recently been introduced for Kartvelian languages, most notably for Georgian (Lobzhanidze et al., 2024), no comparable syntactic resource currently exists for Megrelian. The MLC therefore constitutes the first step toward establishing a UD-compatible treebank for the language.

3. Annotation and Corpus Development

3.1. Data Collection, Transcription and Preparation

To ensure that the corpus reflects contemporary Megrelian usage, new data were collected through extensive fieldwork (2022-2024) rather than relying on archival materials. Following established documentation principles (Austin, 2006; Bowern, 2008), recordings were done across diverse genres and sociocultural contexts, including narratives, rituals, toponyms and everyday discourse. The fieldwork produced 58 hours of high-quality audio and video recordings from speakers representing both major dialect zones and multiple age groups (15-30, 31-45, 46-60, 61+). Each session was accompanied by detailed metadata covering demographic and sociolinguistic variables and governed by informed consent protocols.

The audio and video recordings were transcribed into the IPA using a custom-developed Megrelian Converter (Gersamia and Lobzhanidze, 2022–2025a). The transcribed texts were then imported into FLEx, where each text was reviewed, segmented and morphosyntactically annotated. Every sentence was aligned with free translations into Georgian and English, produced collaboratively by native Megrelian-Georgian bilinguals and speakers proficient in English, who were also responsible for the morphosyntactic annotation of the corpus (see Figure 1).

3.2. Morphological and Lexical Annotation

After translation, each sentence was tokenized and each token was morphologically annotated and accompanied by English glosses. The FLEx environment provided a structured XML framework for encoding this information (see Figure 2), including fields for lemma, morpheme segmentation, grammatical category and gloss. This hierarchical annotation structure is reflected in the relational schema shown in Figure 3. Because FLEx maintains a morpheme-based lexicon, recurring stems and affixes were automatically recognized and linked to existing entries, while new tokens generated provisional lexical items with glosses. This dynamic lexicon, comprising 97,691 tokens and 60,959 types, ensures up-to-date corpus statistics and enables consistent export for further integration into the corpus interface.

As was mentioned above, the morphosyntactic annotation and glossing were carried out by trained linguists with expertise in Megrelian, Georgian and English. To ensure consistency and accuracy, the annotation process was supervised by

```
  <phrase guid="817988fa-0df2-46aa-8fff-92403b3396a6">
    <item type="txt" lang="xmf">- დო თე კაჭიჭი მუჭომ ჯურძენიდ ან მუ ფერიდ?</item>
    <item type="segnum" lang="en">1</item>
    <words>
    <item type="gls" lang="en">What kind of grape was Kachichi? What color was it?</item>
    <item type="gls" lang="xmf-fonipa">dɒ tɛ kʼɑtˇʃʼitˇʃʼi mutˇʃʼɔm ʔurdˇzɛnid ɑn mu pɛrid?</item>
    <item type="gls" lang="ka-Brai">- და ეს კაჭიჭი როგორი ყურძენი იყო ან რა ფერის იყო?</item>
  </phrase>
</phrases>
```

Figure 1: Phrase-level annotation in FLEx showing interlinear structure

```
<word guid="5e2deb9a-323a-402a-8eb4-5fa1d019474a">
  <item type="txt" lang="xmf">კაჭიჭი</item>
  <morphemes>
    <morph type="stem" guid="d7f713e8-e8cf-11d3-9764-00c04f186933">
      <item type="txt" lang="xmf">კაჭიჭ</item>
      <item type="txt" lang="xmf-fonipa"></item>
      <item type="cf" lang="xmf">კაჭიჭ</item>
      <item type="gls" lang="en">Katchitchi</item>
      <item type="msa" lang="en">pn</item>
    </morph>
    <morph type="suffix" guid="d7f713dd-e8cf-11d3-9764-00c04f186933">
      <item type="txt" lang="xmf">-ი</item>
      <item type="txt" lang="xmf-fonipa">-i</item>
      <item type="cf" lang="xmf">-ი</item>
      <item type="hn" lang="xmf">1</item>
      <item type="gls" lang="en">NOM</item>
      <item type="msa" lang="en">n:(Case)</item>
    </morph>
  </morphemes>
  <item type="pos" lang="en">pn</item>
</word>
```

Figure 2: Word-level annotation

the project's Principal Investigator, who conducted a systematic quality control of the annotated material.

The validation procedure involved manual inspection of tokenization, morpheme segmentation, part-of-speech assignments and gloss consistency. Special attention was paid to complex agglutinative verbal forms and clitic constructions, which are particularly prone to segmentation inconsistencies. Identified discrepancies were discussed with annotators and resolved collaboratively and annotation guidelines were refined iteratively to prevent recurring inconsistencies.

3.3. Conversion of FLEx Data into the MLC Database

After tokenization, lemmatization and interlinear glossing were completed in FLEx, each text was exported as Verifiable Generic XML (see Figures 1–2). The exported files were then processed through a custom Python-PHP conversion pipeline. The XML output was reformatted by a Python script and passed to a PHP-based loader, which subsequently executed the following four operations simultaneously: a) opened a UTF-8 connection to the project database; b) parsed the XML hierarchy (interlinear-text -> paragraphs -> phrases -> words -> morphemes) using the xml_load_file() function; c) extracted the tier values for each node and inserted them into the corresponding database tables; and d) preserved all FLEx globally unique identifiers (GUIDs) as relational keys.

The resulting relational schema (see Figure 3) maps the hierarchical structure of the FLEx XML, with each table representing a specific annotation level and connected through the retained GUIDs.

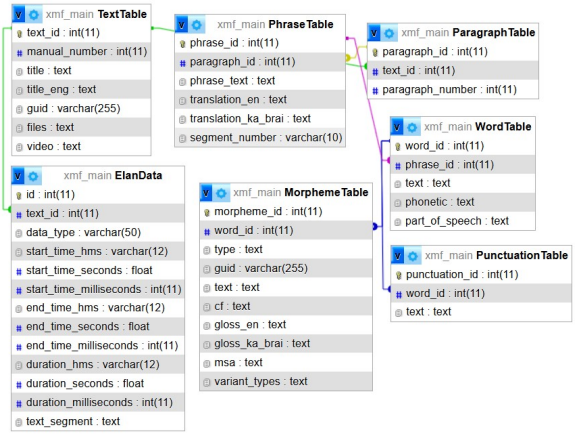

Figure 3: Database schema

3.4. Corpus Interface and Functionality

The corpus search interface consists of two components. The first component represents the linguistically annotated data exported from FLEx and allows users to explore Megrelian texts at three levels:

- **Texts:** Full transcribed texts with aligned Georgian and English translations, audio/video links and clickable morphemes displaying glosses and PoS information.
- **Lines:** A sentence-level view with searchable Megrelian-English parallel data, linked to the corresponding full text contexts.
- **Morphemes:** A bilingual Megrelian-English dictionary view, showing grammatical features, glosses and frequency of occurrences, with direct links to all corpus instances.

The second component provides access to early published Megrelian texts that lack linguistic annotation. This interface supports metadata-based queries through fields such as Title, Author, Publication Date and Publication Place, as well as

full-text searches via the keyword bar (See Figure 4). Users can refine searches by combining metadata filters or select the 'Search all texts (ignore filters)' option to browse the entire collection. Search terms may include exact tokens, phrases or wildcard patterns, providing flexible access to textual data. The preview panel displays the main bibliographic information (Title, Author, Publication Date and Publication Place), allowing users to identify relevant materials before viewing the full text. This component is particularly valuable for locating Megrelian publications and unannotated data, which can serve as reference materials for future annotation and diachronic linguistic analysis.

4. Initial Steps Toward UD Alignment

4.1. Mapping FLEx Annotations to UD

The alignment of Megrelian data with the UD framework was based on the morphosyntactic information exported from FLEx. Each annotated text was exported as verifiable generic XML, which preserves the interlinear structure (interlinear-text -> paragraph -> phrase -> word -> morpheme) and provides detailed information on lexical, morphological and glossing tiers. These fields were systematically mapped to the core UD layers (LEMMA, UPOS, XPOS and FEATS) representing the first step toward UD integration for Megrelian.

The mapping procedure was carried out using a custom Python conversion script, which parsed the XML hierarchy and extracted relevant linguistic information. The lexical form of each token in FLEx was assigned to the lemma field, while the gloss and part of speech tiers were normalized and aligned to the UD's UPOS and XPOS categories. Morphological attributes (e.g., case, number, person, tense and aspect) were encoded in the FEATS column following UD conventions. Particular care was taken in handling agglutinative verbal forms, which required rule-based segmentation to ensure that linguistic information was properly represented in UD format. To illustrate the conversion process, let's consider the complex verbal form ქვდარიბუანქი (kədɑrinuɑnki) 'you (sg) will make for him/her emphatically'. In FLEx, the verb is segmented into multiple morphemes reflecting preverbation, applicative morphology, thematic suffixation, agreement and emphatic marking:

ქვ-დ-ა-რიბ-უ-ან-ქ-ი
kə-d-a-rin-u-an-k-i
AFFMT-PRV-APPL.INDIR-BE-AUG-TS-SUBJ2SG-EMPH
'you (sg) will make for him/her emphatically'

In the resulting CoNLL-U representation, these elements are encoded within the FEATS column as follows:

```
text = ქვდარიბუანქი
1  ქვდარიბუანქი  რიბ  VERB  AFFMT_PRV
_APPL.INDIR_AUG_TS_SUBJ2SG_EMPH
Aspect=Perf|Mood=Ind|Number[subj]=Sing|
Person[io]=3|Person[subj]=2|Subcat=Tran|
Tense=Fut|Voice=Act  _  _  _
Translit=kəd rinu nki
```

Here, the lemma field preserves the lexical root (რიბ (rin), which can be translated as 'do', 'make' or 'exist', depending on the preverb and surrounding morphemes), while derivational and inflectional morphology is represented through UD-compliant feature bundles in the FEATS column.

Given the typological similarities between the Megrelian and Georgian, both belonging to the Kartvelian family, the Georgian UD treebank served as a key reference point during this process. However, due to the lack of an existing tag-set or morphological analyzer for Megrelian, the mapping required defining feature sets manually, directly from the interlinear glosses produced in FLEx. The resulting CoNLL-U files preserve the original morphological detail while introducing a consistent structure suitable for subsequent dependency annotation.

4.2. Initial Annotation Scheme

4.2.1. Tokenization and Word Segmentation

In Megrelian, words are generally delimited by whitespace and punctuation marks. Punctuation symbols are not separated from adjacent words, including in hyphenated compounds such as xɑtɛ-xɑtɛ 'suddenly-suddenly', which constitute a single token. However, the dash is treated as a separate character. Complex punctuation sequences such as ?!, !.. and …, as well as decimal numerals (e.g., 1.2, 0,5), are preserved as single tokens.

Due to Megrelian's rich agglutinative morphology, certain clitic constructions are analyzed as multiword tokens, segmented into individual syntactic words in the following cases: a) Postpositions attached as suffixes to inflected nominals: k'ɔt͡ʃit͡s'k'ɑlɑ 'with a human'; ʔudɛʃɑx 'until the house'; b) Indirect speech particles attached to inflected nominals or verbs: k'ɔt͡ʃkiɑ 'a man, as someone said'; vt͡ʃ'ɑrunkiɑ 'I write, as someone said'.

4.2.2. PoS Tagging

The UD_Megrelian-MLC treebank uses all core Universal POS (UPOS) categories, while the original FLEx tags are preserved at the language-specific XPOS level. Particles (PART) occur both as independent words (e.g., xɔlɔ 'too') and as clitics attached to nominal or verbal paradigms (e.g., -va(r)- 'no', -ɔ interrogative particle). Independent particles receive the PART tag, while cl-

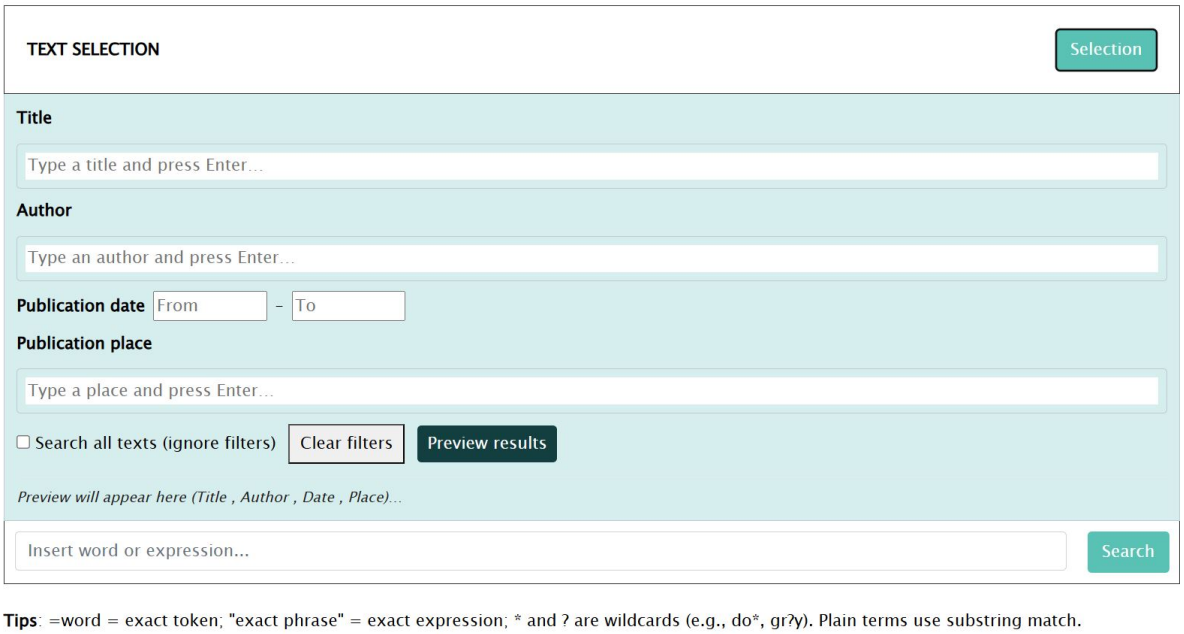

Figure 4: Text search interface.

itics are marked with PartType=Emp, etc. Concerning auxiliaries (AUX), the copula i?in 'to be' expresses existence, identity or predication, appearing in various forms to agree with person, number and tense. As about verbal derivatives, verbal nouns and adjectives are assigned VERB with VerbForm=Vnoun or VerbForm=Part respectively.

4.2.3. Morphological Features

The UD_Megrelian-MLC treebank uses all major lexical FEATS defined by the UD framework (e.g., PronType, NumType, Poss, AdpType, NameType, PartType, etc.).

- **Nominal Features**: Megrelian nouns inflect for Number (singular/plural) and Case (nominative, ergative, dative, genitive, allative, ablative, benefactive, instrumental and essive); Adjectives generally agree with their head nouns in Case and Number and may also inflect for Degree (diminutive, equative, superlative); Numerals and pronouns follow similar patterns, with personal, demonstrative and possessive pronouns marking Person (1, 2, 3).
- **Verbal Features**: Verbs inflect for Tense (present, imperfect, future, aorist, perfect, pluperfect), Aspect (imperfective, perfective) and Mood (indicative, subjunctive, conditional); The system is organized into TAM series, which determine case-marking and agreement relationships between agent and patient through preverbs, version markers and thematic suffixes (Gersamia, 2022); Verbs agree with their subject and objects (direct and/or indirect) in Person and Number, exhibiting split-ergative alignment typical of Kartvelian languages; Nonfinite forms include participles (VerbForm=Part) and masdars (VerbForm=Vnoun).

5. Conclusions and Future Work

This paper presented the MLC, the first publicly available, linguistically annotated resource for the Megrelian language, an endangered member of the Kartvelian family. The corpus integrates spoken and written data collected through recent fieldwork, processed and annotated in FLEx. The UD alignment, covering the lemma, UPOS, XPOS and FEATS layers and annotation of the initial 150 utterances can be considered as the beginning for the development of a full Megrelian UD Treebank. Future work will focus on expanding the corpus with new transcriptions and texts, extending syntactic annotation to include dependency relations (DEPREL) and enhancing automatic processing through morphological tagging and parsing tools. Ultimately, the MLC aims to serve both the research community and native speakers, supporting the documentation, revitalization and computational study of Megrelian within the broader context of multilingual NLP.

6. Acknowledgements

This work was supported by Shota Rustaveli National Science Foundation of Georgia (SRNSFG) [grant number: FR-21-993-3, 2021-2025]. Irina Lobzhanidze also acknowledges support from COST Action CA21167. The authors express their gratitude to the three anonymous reviewers for the many helpful comments and suggestions.

7. Limitations

While the current work establishes a foundation for UD alignment in Megrelian, some limitations remain. The pilot dataset is relatively small and currently limited to the lemma, UPOS, XPOS and FEATS layers. Further expansion will require manual review to identify potential inconsistencies in morphological alignment and to extend coverage to DEPREL.

8. Ethics Statement

All fieldwork and data collection activities for the Megrelian Language Corpus (MLC) were conducted in accordance with the ethical guidelines of Ilia State University and ethical approval was obtained prior to the start of the project. Each participant provided written informed consent in both Georgian and English, with separate permissions for (i) recording, (ii) publication of audio or video materials and (iii) publication of textual data. Participation was entirely voluntary and speakers were informed of their right to withdraw from the project at any time.

Since the data include publicly available materials (video recordings), full anonymization was not possible. However, the corpus release (version 1.0) excludes sensitive personal information beyond what is already publicly accessible. The annotated corpus does not contain any sensitive, confidential or offensive content. All texts and recordings are based on data collected during linguistic expeditions and on openly available materials for which explicit consent was obtained.

Annotation and translation were performed by four trained bilingual Megrelian-Georgian annotators, who were involved in the realization of the project under expert supervision to ensure accuracy and consistency.

We have no conflicts of interest to disclose.

9. Bibliographical References

Peter K. Austin. 2006. Data and language documentation. In Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, editors, *Essentials of Language Documentation*, pages 87–113. Mouton de Gruyter, Berlin and New York.

Claire Bowern. 2008. *Linguistic Fieldwork: A Practical Guide*. Palgrave Macmillan, New York.

Arnold Chikobava. 1936. *čanuris gramatikuli analizi: tek stebit a da ganmartebebit [Grammatical analysis of Zan: accompanied by texts and explanations]*. enisa, istoriisa da materialuri kulturis institutis (enimkis) moambe [Bulletin of the Institute of Language, History, and Material Culture (ENIMKI)], Tbilisi.

Nana Danelia and Inga Dundua. 2006. megruli enis prak tikuli kursi [practical course of the megrelian language]. In Rusudan Amirejibi-Mullen, Nana Danelia, and Inga Dundua, editors, *kolxuri (megrul-lazuri) ena [Colchian (Megrelian–Laz) language]*, pages 175–339. Universal, Tbilisi.

Givi Eliava. 1997. *megrul-k art uli lek sikoni [Megrelian–Georgian Dictionary]*. Intellect, Tbilisi.

Rusudan Gersamia. 2020. *Linguistic Representation of Space and Motion: An Analysis of Laz Language Data*. Ilia State University, Tbilisi.

Rusudan Gersamia. 2022. The morphonemics of verbal prefixes in megrelian. In Nora Boneh, Daniel Harbour, Ora Matushansky, and Isabelle Roy, editors, *Construire sur les décombres de Babel: Building on Babel's Rubble*, pages 37–59. Presses Universitaires de Vincennes (PUV), Paris 8.

Johann Anton Güldenstädt. 1787–1791. *Reisen durch Russland und im Caucasischen Gebirge [Travels through Russia and in the Caucasus Mountains]*. Kaiserliche Akademie der Wissenschaften, St. Petersburg. Volumes 1–2.

Otar Kajaia. 2000–2009. *megrul-k art uli lek sikoni [Megrelian–Georgian Dictionary]*. Nekeri, Tbilisi. Volumes 1–4.

Guram Kartozia. 2008. *megruli da lazuri tek'stebi [Megrelian and Laz texts]*. Nekeri, Tbilisi.

Guram Kartozia, Rusudan Gersamia, Maia Lomia, and Taia Tskhadaia. 2010. *megrulis lingvisturi analizi [Linguistic analysis of Megrelian]*. Meridiani, Tbilisi.

Bezhan Kilanava. 2010. *900 megruli sitqva [900 Megrelian Words]*. Intellect, Tbilisi.

Iosef Kipshidze. 1914. *Grammatika mingrel'skogo (iverskogo) yazyka s khrestomatiey i slovarem [Grammar of the Mingrelian (Iverian) language with chrestomathy and dictionary]*. Tipografiya

Imperatorskoy Akademii Nauk [Printing House of the Imperial Academy of Sciences], Saint Petersburg.

Julius von Klaproth. 2012. *Reise in den Kaukasus und nach Georgien: unternommen in den Jahren 1807 und 1808 [Journey to the Caucasus and to Georgia: undertaken in the years 1807 and 1808]*. Nabu Press [Hallisches Waisenhaus], Charleston, SC [Halle]. Originally published 1812–1814.

Alio Kobalia. 2010–2020. *k art ul–megruli lek sikoni [Georgian–Megrelian Dictionary]*. Artanuji, Tbilisi.

Arcangelo Lamberti. 1654. *Relatione della Colchide hoggi detta Mengrella, nella quale si tratta dell'origine, costumi, e cose naturali di quei paesi [Description of Colchis, today called Mingrelia, discussing the origins, customs, and natural features of those lands]*. Camillo Caualli, Napoli.

Irina Lobzhanidze, Erekle Magradze, Svetlana Berikashvili, Anzor Gozalishvili, and Tamar Jalaghonia. 2024. Building a Universal Dependencies treebank for Georgian. In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*, pages 40–45, Hamburg,Germany. Association for Computational Linguistics.

Givi Machavariani. 2002. *k art velur enat a šedarebit i gramatika (lek c iebis kursi) [Comparative grammar of the Kartvelian languages: Course of lectures]*. saxelmcip o universitetis gamomc emloba [State University Press], Tbilisi.

Aleksandre Tsagareli. 1880. *Mingrel'skie ėtiudy, vypusk I [Megrelian studies, Part I]*. Tipografiya Imperatorskoy Akademii Nauk [Printing House of the Imperial Academy of Sciences], Saint Petersburg.

10. Language Resource References

Rusudan Gersamia and Irina Lobzhanidze. 2022–2025a. *The Megrelian Converter (MC)*. Ilia State University. Ilia State University. Accessed 20 July 2025.

Rusudan Gersamia and Irina Lobzhanidze. 2022–2025b. *The Megrelian Language Corpus (MLC)*. Ilia State University. Ilia State University. Accessed 20 July 2025.

Jost Gippert, Javier Martínez, Agnes Korn, and Roland Mittmann. 2016. Armazi corpus project. Accessed 20 July 2025.

Jost Gippert and Paul Meurer and Manana Tandashvili, M. 2011–2025. *The Georgian National Corpus (GNC)*. Tbilisi State University. Georgian National Corpus Project. Accessed 20 July 2025.

Max Planck Institute for Psycholinguistics. 2025. *ELAN (Version 7.0) [Computer software]*. The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

SIL International. 2025. *Fieldworks Language Explorer (FLEx), Version 9.1*. SIL International. Accessed 20 July 2025.

UNESCO. 2021. Atlas of the world's languages in danger. Online edition accessed 20 July 2025.