# From Interlinear Annotation toward a UD Treebank for Megrelian language

**Irina Lobzhanidze**
Ilia State University, Tbilisi, Georgia;
irina_lobzhanidze@iliauni.edu.ge

**Rusudan Gersamia**
Ilia State University, Tbilisi, Georgia;
rgersamia@iliauni.edu.ge

***Relevant UniDive working groups:*** WG1, WG2

## 1 Introduction

Low-resourced and endangered languages remain significantly underrepresented in computational linguistics due to the scarcity of annotated corpora, trained annotators and interoperable annotation frameworks. Megrelian (ISO 639-3: xmf), an increasingly endangered Kartvelian language (UNESCO, 2021) spoken in Western Georgia, represents these challenges, combining rich agglutinative morphology with limited digital resources and predominantly oral usage. The lack of standardized orthography and computational tools further complicates linguistic analysis and resource development. This abstract presents the Megrelian Language Corpus (MLC) and its ongoing transformation into a Universal Dependencies (UD)-compatible resource, aiming to support both language documentation and multilingual NLP research.

The MLC comprises both spoken and written data collected through fieldwork conducted between 2022 and 2024 in the Samegrelo-Zemo Svaneti region. Audio and video recordings were first manually transcribed by Megrelian-speaking annotators using the Megrelian script based on the Georgian alphabet. Given the lack of fully standardized orthographic conventions, the resulting transcriptions were normalized to ensure annotation consistency and subsequently converted to the International Phonetic Alphabet (IPA) using the Megrelian Converter and manually verified prior to morphosyntactic annotation in FLEx (SIL International, 2025). Both video recordings and FLEx-annotated materials were processed and time-aligned in ELAN (Max Planck Institute for Psycholinguistics, 2025). The annotated data are further converted into a Universal Dependencies (UD) (Nivre et al., 2017) representation by mapping FLEx-based interlinear annotations to UD layers (LEMMA, UPOS, XPOS, FEATS) and exporting them in CoNLL-U format.

## 2 The Megrelian Language Corpus

The MLC (Gersamia and Lobzhanidze, 2022–2025b) available at https://xmf.iliauni.edu.ge/ aims to document present day usage of this Kartvelian language. The dataset comprises approximately 97000 word tokens and 60000 word types as well as around 58 hours of audio and video recordings collected from speakers representing different dialect zones and age groups. The collected material includes narratives, sociocultural discourse and questionnaire-based interviews, accompanied by detailed metadata on speakers' linguistic and demographic profiles.

Following data collection, the normalized Megrelian transcriptions were converted into IPA using the Megrelian Converter (Gersamia and Lobzhanidze, 2022–2025a) and processed in FLEx. Each sentence was presented in its transcribed form and aligned with free translations into Georgian and English, produced collaboratively by bilingual trained annotators. The annotation workflow involved tokenization and morphological analysis in accordance with the Leipzig Glossing Rules (Max Planck Institute for Evolutionary Anthropology, 2015). Each token was segmented into morphs[1] and enriched with lexical, grammatical and glossing information within FLEx's structured XMLs. The interlinear annotation further enabled the automatic generation of a morph-based lexicon by linking recurring stems and affixes across annotated materials to their corresponding lexical entries in FLEx. The annotated data were then exported as XML and transformed into a relational database through a custom processing pipeline, preserving the hierarchical structure of texts, phrases, words and morphs.

[1] Here, the term morph refers to surface segmentation units in FLEx interlinear annotation.

On this basis, a corpus with an integrated dictionary was developed. Its provides a multi-level interface designed to support both linguistic analysis and user-friendly access to annotated data. The corpus allows analysis at the level of full texts, individual sentences and morphs. Text-level access presents transcribed Megrelian narratives together with sentence-aligned Georgian and English translations, morph-level glossing and morphosyntactic annotation, as well as linked audio and video recordings. Sentence-level search supports retrieval of parallel Megrelian-English data with direct links to full texts. At the morph level, the corpus functions as a bilingual Megrelian-English lexicon, providing glosses, grammatical features and frequency information, with links to all occurrences in the corpus. The interface also supports metadata-based queries and Key Word in Context (KWIC) format search within unannotated texts published between 1880 and 2012.

## 3 From Interlinear Annotation to UD

The alignment of the MLC with the UD framework is based on the morphosyntactic information encoded in FLEx interlinear annotations. Each annotated text was exported as Verifiable Generic XML, preserving the hierarchical structure of interlinear glossed text (text -> paragraph -> phrase -> word -> morph) and providing access to lexical, morphological and glossing tiers. These layers are systematically mapped to the core UD representation, including LEMMA, UPOS, XPOS and FEATS.

The mapping process was implemented through a custom Python script that parses the XML structure and extracts relevant linguistic information. The lexical entry associated with each token in FLEx is assigned to the lemma field, while PoS tags are normalized and aligned with UD UPOS tags, with original FLEx categories preserved as XPOS. Morphological features such as case, number, tense, aspect, mood and agreement are encoded in the FEATS column following UD conventions. For example, in FLEx, the verb is segmented into multiple morphs representing preverbation, applicative morphology, thematic suffixation, agreement, etc.:

დო-ჭახ-უნ-დ-ეს
*dɔ-t͡ʃʼax-un-d-ɛs*
PRV-PRESS-TS-IMPF-SUBJ3PL
'they used to press'

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

```
1 დოჭახუნდეს ჭახ VERB PRV_TS_IMPF
_SUBJ3PL Mood=Ind|Number[subj]=Plur|
Person[subj]=3|Subcat=Intr|Tense=Imp|Voice=Act
_ _ _ Translit=dɔt͡ʃʼɑxundɛs
```

Particular challenges arise from the agglutinative nature of Megrelian, especially in the treatment of complex verbal forms, which contain multiple prefixes and suffixes encoding agreement, applicative relations, tense-aspect-mood (TAM) series and discourse-related markers. These forms required careful segmentation, rule-based mapping and manual validation to be accurately represented within the UD framework.

In addition, specific decisions were required regarding tokenization and segmentation, particularly in the treatment of cliticized or univerbated forms represented in UD by means of multiword tokens (MWTs). In such cases, a single orthographic form corresponds to multiple syntactic words in the CoNLL-U representation, as follows:

```
1-2 ვაუღ _ _ _ _ _ _ _
SpaceAfter=No
1 ვა PART NEG PartType=Neg _ _ _
Translit=vɑ
2 უღ VERB APPL.OBJ_STAT_SUBJ3
Mood=Ind|Number[subj]=Sing|Person[obj]=3|
Person[subj]=3|Subcat=Tran|Tense=Pres|
Voice=Act _ _ _ Translit=uɣ
```

The resulting UD_Megrelian-MLC dataset currently lacks syntactic annotation layers (HEAD, DEPREL, and DEPS), but preserves rich language-specific morphological information.

## 4 Ethical Considerations

All fieldwork and data collection activities for the 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. All participants provided written informed consent, including separate permissions for recording and for the publication of textual, audio and video materials. Participation was voluntary and participants were informed of their right to withdraw from

the project at any time.

Since the MLC includes publicly accessible video recordings, complete anonymization was not possible. However, the released version of the MLC excludes sensitive personal information beyond what is already publicly available.

## 5 Conclusion

This study introduced the MLC, the first publicly accessible, linguistically annotated corpus with an integrated dictionary for Megrelian. The corpus combines spoken and written data collected through recent fieldwork and annotated using FLEx and ELAN. The current UD alignment includes the LEMMA, UPOS, XPOS and FEATS layers and currently covers an initial subset of 150 utterances from the MLC represented in the UD repository for Megrelian.

Future work will focus on expanding the corpus with additional texts and extending the annotation to cover syntactic relations (DEPREL). The UD-compatible representation is intended as an additional interoperable annotation layer integrated with the broader MLC infrastructure, which preserves aligned audio/video data, interlinear glossing, lexical resources and multilayer morphosyntactic annotation. In the long term, the MLC is intended to serve both researchers and native speakers by supporting language documentation and computational analysis.

## 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: Universality, diversity and idiosyncrasy in language technology (UniDive). The authors thank Prof. Dr. Dan Zeman for his valuable guidance, the three anonymous reviewers for their helpful comments and suggestions and the organizing committee of the 4th UniDive General Meeting.

## References

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

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

Max Planck Institute for Evolutionary Anthropology. 2015. Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses. Available at https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf.

Max Planck Institute for Psycholinguistics. 2025. ELAN (Version 7.0) [Computer software]. 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. Accessed 20 July 2025.

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