LREC 2026

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) @ LREC 2026

Workshop Proceedings

Editors
Antonios Anastasopoulos, Stella Markantonatou, Angela Ralli, Marcos Zampieri, Stavros Bompolas, Vivian Stamou

16 May 2026

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) @ LREC 2026

ISBN 978-2-493814-60-9

Preface

The proceedings contains the 34 papers accepted for presentation at the Workshop on Dialects in NLP — A Resource Perspective (DialRes-LREC2026, https://dialres.github.io/dialres/index.html), co-located with the Fifteenth Biennial Language Resources and Evaluation Conference (LREC). The conference was held at the Palau de Congressos de Palma in Palma de Mallorca, Spain, from 11 to 16 May 2026, with the DialRes workshop taking place on 16 May 2026.

For this first edition of DialRes, we received 43 submissions. We are deeply grateful to the dialect studies and NLP communities for embracing DialRes as a venue for presenting their work. This strong response highlights the growing interest in dialectal research and confirms the need for a dedicated forum focusing on the development and use of dialectal resources.

The accepted papers cover both living and historical dialects from nearly all continents. Contributions on historical dialects include studies on English, German, Heptanesian Greek, Old Irish, and Transcarpathian varieties. Research on living dialects spans a wide range of linguistic contexts, including Arabic, Aromanian, Bangla, Basque, Formosan languages, German, Italian, Kurdish, Modern Greek, and Slavic varieties (such as Pomak and Ukrainian), as well as Wancho.

The contributions address a broad spectrum of topics, ranging from the collection and development of oral and written dialectal resources—such as corpora, treebanks, benchmarks, and specialized databases—to their application in dialectometry and dialect classification. Many papers employ state-of-the-art methodologies, including Automatic Speech Recognition (ASR), neural parsing techniques, and Large Language Models (LLMs).

We would also like to express our gratitude to the Archimedes Research Unit of the Athena Research Center and the UniDive COST Action for endorsing the workshop. In particular, the Archimedes Research Unit generously supported the participation of three of the organizers.

Last but not least, we extend our thanks to the members of the DialRes Program Committee for their contributions to the successful organization of the workshop.

The DialRes Workshop Organizers:

Antonis Anastasopoulos
Stella Markantonatou
Angela Ralli
Marcos Zampieri
Stavros Bompolas
Vivian Stamou

Table of Contents

Pontic Greek in the Caucasus: an online corpus

Svetlana Berikashvili,[1] **Stavros Skopeteas**[2]
Ilia State University,[1] University of Göttingen[2]
Kakutsa Cholokashvili Ave 3/5, Tbilisi 0162,[1] Kate-Hamburger-Weg 3, Göttingen 37073[2]
svetlana.berikashvili@iliauni.edu.ge,[1] stavros.skopeteas@uni-goettingen.de[2]

Abstract

This article presents a multi-media corpus of Pontic Greek as spoken in the Caucasus (Georgia). The corpus covers three major stages reflecting different sociolinguistic settings: (a) rural communities in the original settlements in Georgia; (b) internal migration to urban centers within Georgia; (c) external migration to Greece. The dataset comprises 373 audio recordings (total duration 7h 26m; total word count: 43,073) and is released as an open-access resource. It includes audio files (WAV) and annotations (XML). Annotations provide orthographical transcription, morphemic transcription and morpheme-by-morpheme and sentence-by-sentence translations in English (Toolbox). Transcriptions are time-aligned with the audio files (ELAN) and can be queried online (ANNIS). The corpus is intended for linguists working on dialectology, language contact, and language change, as well as for a broader audience interested in the history and practices of Pontic Greek communities. Pontic Greek in the Caucasus offers a unique opportunity to investigate contact between Greek and Indo-European (Russian) as well as Non-Indo-European languages (Georgian, Turkish) in different sociolinguistic settings.

Keywords: Pontic Greek, corpus resource, corpus design, corpus annotations, endangered dialects

1. Introduction

The development of dialectal oral and textual resources is essential for our knowledge of linguistic heritage and language variation. This paper presents a dataset designed to capture variation in Pontic Greek, as spoken by the Pontic speakers of Georgia, whose ancestors migrated from Anatolia in several waves beginning in the nineteenth century. Pontic Greek in the Caucasus remains underrepresented in research on Greek dialects and, in light of the migration processes of the past three decades, has become severely endangered. Its systematic documentation in online resources is therefore an urgent priority.

The data for the present corpus were collected from Pontic speakers in Georgia as well as from speakers who have migrated to Greece. The dataset comprises a morphologically annotated corpus of narrative texts and additional interviews (with a total recording time of 7h 26m). Beyond their linguistic relevance, these materials also provide insights into cultural practices. The resource is available both as an online corpus at `https://spw.uni-goettingen.de/projects/xtyp/PNT.html` and as downloadable files through the TLA (Kotanidi et al., 2019).

The relevant background of Greek populations in the area is outlined in Section 2. Section 3 reviews related resources on Greek dialects. Section 4 presents the methodology of the data collection. Section 5 outlines the annotations of the corpus and Section 6 the online resources. Section 7 illustrates the use of the corpus, and Section 8 concludes.

2. Greek Populations in the Caucasus

The ethnic Greek population in the Caucasus consists of two linguistic groups: (a) Urum speakers, whose language is closely related to Anatolian varieties of Turkish, and (b) Pontic speakers, who use the Pontic dialect of Greek; see Sideri (2006) and Höfler (2020) on the interplay between language and ethnic identity. Both groups migrated to Georgia from the Ottoman Empire in several waves. Approximately 800 families had already settled in eastern Georgia in the eighteenth century (Kaukhchishvili, 1946); Greek villages in Armenia were likewise established in the eighteenth century as a result of labor migration (Hodgson, 2010). Major migration waves to Georgia followed the end of the Russo-Turkish war of 1828-1829 (42,000 people), the Crimean war 1853-1856 (36,000 people), and a further wave took place between 1878 and 1882 (17,100 people) (Xanthopoulou-Kyriakou, 1997).

Pontic Greeks originated from various areas of northeastern Anatolia, including Kars, Erzurum, Trabzon, and Gümüşhane (see triangles in Figure 1). They settled in numerous villages and towns across Georgia. The earliest documented settlement (since 1829) is Ts'ints'q'aro (Kaukhchishvili, 1942). Other villages are located in Tetritskaro region (e.g., Big Iraga, Small Iraga, Jigrasheni, etc.), in Dmanisi (Gora, Sakire), in Borjomi (Tsikhisdjvari), along the Black Sea coast in Adjara (Dagva, Kvirike, Akhalsheni) as well as in Abkhazia (Aleksandrovka, Mikhailovka). Most villages in the Ts'alka region (south of Tbilisi) are Urum-speaking, although Pontic Greek is also reported in some settlements

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 230–237
16 May 2026.

(Santa, Neokharaba). In northern Armenia, the Greek villages are concentrated around Hankavan.

Official records only report ethnic affiliation, but do not distinguish groups by language. The ethnic Greek population in Georgia reached its peak in 1989, with 100,324 individuals. Following large-scale migration to Greece in the 1990s, the number declined to 15,100 in 2002 and further to 5,500 in 2014 (see Loladze, 2021 for the dynamics of the Greek population in Georgia from 1800 to 2014).

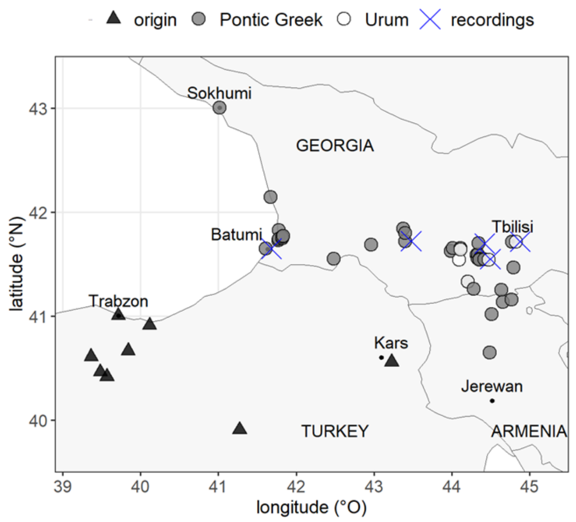

Figure 1: Greek populations in the Caucasus

Since Pontic Greek in the Caucasus originates from different locations of Anatolia, it is dialectally heterogeneous. However, the extent of koineization processes in the new settlements remains to be investigated. A common characteristic of these varieties is their development in a multilingual environment, shaped by sustained contact with Russian (since the nineteenth century), with Georgian (particularly in urban centers in recent decades), as well as with Standard Modern Greek, especially following renewed transnational contacts and labor migration since the 1990s. The corpus presented in this article has been designed to capture exactly these processes across different types of settlements.

3. Related Resources

Most available research and resources on Pontic Greek concern the varieties spoken in Anatolia, based either on data collected in Turkey or from emigrant speakers in Greece. The grammatical and historical dimensions of the dialect have been examined in comprehensive grammars, grammatical sketches and related studies (Papadopoulos, 1955; Oikonomidis, 1958; Mackridge, 1987; Tombaidis, 1988; Drettas, 1997; Revithiadou and Spyropoulos, 2012; Sitaridou, 2013, among many others).

Accessible digital databases containing primary Pontic data remain limited. A major resource is the AMiGre database (Pontus, Cappadocia, Aivali: in search of Asia Minor Greek),[1] which provides both raw and processed material from three dialects. The database includes an oral corpus of approx. 180h (with a multimodal sub-corpus of around 45h), a written corpus, and an electronic tri-dialectal dictionary.[2] Part of the corpus (15h per dialect) is annotated (Karanikolas et al., 2015).

Additional resources provide materials for special purposes. A dataset designed for NLP-based dialect identification includes Pontic (282,000 words) alongside other Greek dialects (Chatzikiriakidis et al., 2025). Pontic data are also available within the WOWA (Word Order in Western Asia) corpus, a spoken-language resource aimed at investigating areal effects in word order variation (Schreiber, 2021). Ethnographically collected Pontic data (without linguistic annotation) form part of the VLACH, Vanishing Languages and Cultural Heritage Collection. [3] Audio recordings representing various dialectal varieties, including Pontic (18 sound-files, 2 of which represent Pontic Greek in Georgia), are accessible through the Ulster University repository. [4] Large resources of written and spoken materials have also been used for the creation of the online atlas of Cappadocian Greek, DiCaDLand (Digitization of the Cappadocian Dialectal Landscape), which, however, does not include Pontic Greek (Melissaropoulou et al., 2022).[5]

Pontic Greek as spoken in the Caucasus remains comparativelly under-resourced and less systematically investigated. Existing grammatical descriptions include Lazarev (1950), Eloeva (1997), Mikaberidze and Shakhpazidi (2000), and Berikashvili (2017) on Pontic Greek in Georgia as well as Hodgson (2010) on Pontic Greek in Armenia. However, comprehensive and digitally accessible resources for these varieties are still lacking.

4. Data Collection

4.1. Speakers

Participants were informed about the aims of the data collection and provided written informed consent for their data to be used in scientific research and deposited in online repositories. The speaker

[1] https://amigre.upatras.gr/
[2] http://amigredb.philology.upatras.gr
[3] https://www.oeaw.ac.at/vlach/collections/greek-varieties/pontic-greek
[4] https://www.ulster.ac.uk/research/topic/modern-languages-linguistics/quality/research-projects/investigating-variation-and-change
[5] http://cappadocian.upatras.gr/en/node/10

sample represents three (idealized) sociolinguistic settings:

- Stage A: HOMELAND: villages traditionally populated by Pontic Greek speakers: Manglisi, Tetritskaro, Orkhevi, Santa, Tsikhisjvari (see "recordings" in Figure 1).
- Stage B: INTERNAL MIGRATION: urban centers of Georgia, with Pontic Greek speakers living in strongly multilingual/multiethnic contexts: Tbilisi.
- Stage C: EXTERNAL MIGRATION: Pontic Greek speakers from Georgia who migrated to Greece: Thessaloniki (speakers originating from different places in Georgia).

Demographic information for each stage is summarized in Table 1. The smaller sample in stage B is due to the difficulty of identifying Pontic Greek speakers in urban centers such as Tbilisi.

stage	*n*	F	M	year of birth		
				min	max	mean
A	18	12	6	1928	1990	1958
B	8	4	4	1925	1989	1950
C	16	7	9	1939	1994	1975

Table 1: Speaker samples (total: *n*=42)

The metadata include information on the location and date of the recordings, as well as speaker-related variables: (a) gender and year of birth and (b) self-reported frequency of using Pontic Greek, Russian, Georgian, and Standard Modern Greek in the communication with family members and neighbors. Figure 2 presents aggregated results of self-assessed language use, which are informative for the vitality of the dialect in the sociolinguistic settings at issue. Speakers assess (on a 1-5 Likert scale) the frequency of use of Pontic Greek and the corresponding dominant language (Russian in stages A/B; Standard Modern Greek in stage C). Stage A speakers show balanced use of Pontic vs. Russian (dominant lg.), whereas the use of the dominant language (Russian or Greek) increases in stages B and C. In the latter stages, Pontic Greek is primarily used within the family and more frequently with the older generation (parents) than with children. Across stages, the use of Pontic Greek increases with social distance (more Pontic within the family) and with the age of the addressee (more Pontic when communicating with the older individuals, with the age effect being stronger within the family).

These self-estimated data confirm the rationale of sampling speakers across three stages, reflecting distinct sociolinguistic situations that influence language use.

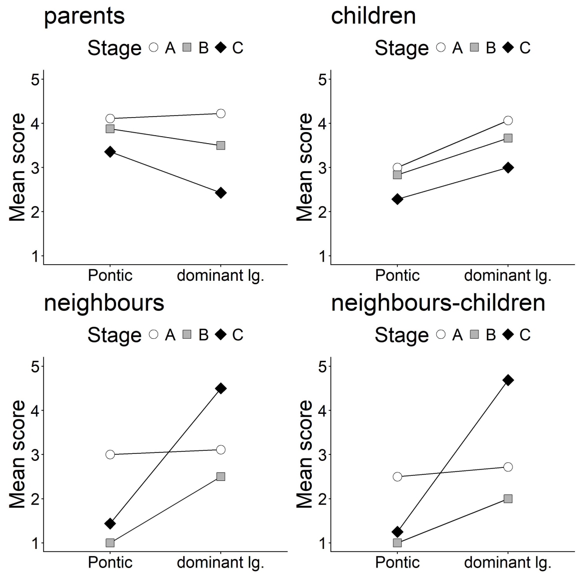

Figure 2: Aggregated self-estimation scores of the language use with different addressees

4.2. Texts

This corpus was designed specifically to study variation between speakers in the stages introduced in §4.1. To this end, all participants were instructed to produce the same eight narratives on the following topics: ANCESTORS, FAMILY, VILLAGE, CULTURE, PEOPLE, MARRIAGE, FEAST, and LANGUAGE. Four of the prompts are illustrated below; the remaining prompts are available in the online documentation of the resource:[6]

- ANCESTORS: How did your ancestors come to Georgia?
- COMPARISON: Please tell us how your people are different from the other people in the village/city?
- MARRIAGE: Please tell us how your people celebrate an engagement/marriage and what is the difference to the way other people in this village/city celebrate an engagement/marriage.
- CULTURE: 'Please tell me a fairy tale or a poem in your native language. If you do not know any fairy tale or poem, please tell me what you consider most important in the culture of your people.'

Consequently, this corpus offers a dataset for special purposes, which may not have the breadth to serve as the basis for a comprehensive grammatical description. For instance, the restriction to narrative texts already implies that certain speech acts,

[6] https://spw.uni-goettingen.de/projects/xtyp/PNT.html

such as directives, are only marginally represented. This corpus is suitable for investigating variation between individual speakers or speaker groups on the basis of linguistic phenomena that recur in narratives, such as the realization of specific phonemes, intonation patterns, frequent morphosyntactic categories such as case marking and definiteness, word order, etc. From an ethnographic perspective, the corpus provides information for the specific cultural practices described in the eight narratives.

Beyond the main text collection (Kotanidi et al., 2019), the online resource also includes additional texts collected during the project (Berikashvili, 2019; Berikashvili and Skopeteas, 2019), that are not discussed here in detail.

4.3. Data Collection Procedure

The interviews were conducted in non-laboratory environments. Speakers were instructed in Pontic Greek by a native speaker of the dialect. Before the questions on the individual narratives, speakers were received the general instruction:

'Please answer the following questions spontaneously. Speak naturally, as if you are speaking to a friend. It does not matter if you are unsure about details, just give a natural response.'

Recordings were made using an Olympus LS-11 recorder with built-in microphones, at a sampling rate of 44.1 kHz. Audio files were saved in .wav format.

5. Annotations

5.1. Data Processing

The recordings were transcribed and annotated by domain experts, combining native-speaker competence and linguistic expertise. Transcriptions were made by a speaker with competence in Pontic Greek, Russian and Georgian. Transcriptions were subsequently morphologically analyzed and glossed by a linguist with expertise in all related languages in Toolbox (SIL, 2015). Toolbox annotations were then exported to ELAN (Sloetjes and Wittenburg, 2008), where sentence boundaries were time-aligned with the audio files. The annotation is expert-based and was not designed for inter-annotator agreement, which is difficult to implement for endangered varieties with limited expert availability.

Project materials include soundfiles (.wav) and annotations in xml format (.eaf files, ELAN). All files follow a unified file-naming convention using the template: language-collection-task-additional-speaker. For example, the filename PNT-TXT-AN-00000-A03.wav is interpreted as shown in Table 2.

PNT	*Language identifier* for 'Pontic Greek' (ISO-639-3)
TXT	*Subcollection identifier* TXT= subcollection 'texts'
AN	*Task identifier* AN= 'ANCESTORS'
00000-	*Additional info* no information in this example
A03	*Speaker identifier* speaker 3, sample A (HOMELAND)

Table 2: File-naming template

5.2. Annotation Tiers

The corpus has a multi-layer annotation structure, including transcriptions, morphological segmentation, glossing, POS tagging, and translation. The annotation tiers (in Toolbox and ELAN files) are listed in Table 3. (The tiers marked with the suffix *_a* represent the original annotations, which are subsequently revised in the corresponding primary tiers.)

name	parent	content
tx	-	*text:* revised transcriptions
tx_a	ref	*text_associated:* original transcriptions
mb	tx	*morphemic boundaries:* transcription enriched with morphemic boundaries
ge	mb	*gloss-English:* revised morpheme-by-morpheme translation (English)
ge_a	mb	*gloss-English_ associated:* original word-by-word translation (English)
ps	mb	*part of speech:* part of speech classification
ft	ft	*free translation:* sentence-by-sentence translation (English)
nt	ref	*notes:* free comments

Table 3: Annotation tiers

5.3. Transcriptions

Text transcriptions ('tx' tier) were based on orthographic conventions that are generally close to a broad phonological transcription. Greek words were transcribed using a system based on the phonemic transcription with IPA-symbols, as proposed by Drettas (1997); see (1). The symbol *-ä* represents the phonemic value [æ], appearing in cases of sinezesis (/ia/ > /æ/) as in (1b) or foreign words, as in (1c). A correspondence table (orthographic symbols to phonemic values) as well as further transcription conventions are listed in the corpus website.[7]

(1) a. *érϑan* 'come:PFV.PST:3.PL'
b. *ospítä* 'house:N.PL.NGEN'
c. *tätǻ-m* 'father-CL.1.SG:GEN'

Embedded Russian elements are transcribed according to the BGN/PCGN romanization standard; see (2a). Georgian elements are rendered following the Georgian national system of romanization (Apridonidze and Chkhaidze, 2002); see (2b). When elements of Georgian/Russian origin are morphologically integrated into Pontic Greek (e.g., by inflection), they are transcribed with the orthographic conventions adopted for Pontic Greek (in order to avoid switching orthographies within a word); see (2c). Word stress is indicated in all words by an acute accent (´).

(2) a. *prinimát'* 'take:IPFV.INF'
b. *ch'ianch'véla* 'ant:SG.NOM'
c. *chixirtmás* 'Chikhirtma:F.PL.NGEN' (Georgian chicken soup) Georgian romanization: *chikhirtma*

5.4. Morphological Annotations

Morphological annotations comprise morphological segmentation ('mb' tier) and morpheme-by-morpheme translations ('ge' tier). Morphological annotations were based on The Leipzig Glossing Rules [8] – with addition of abbreviations from Eurotyp (see König et al., 1993).

As a general practice for highly inflectional languages, inflectional forms of Pontic Greek were not segmented with morphemic boundaries (at the 'mb' tier), but morphemes are translated in the 'ge' tier according to category-specific templates, as illustrated in (3) for nominals (substantives, adjectives, pronouns, adjectival participles).

The template for verbs is presented in (4). By convention, unmarked categories (active, indicative, present, finite) are not represented in glossing.

[7]Abbreviations: 1= 1st person; 3 = 3rd person; ACC = accusative; CL = possessive clitic; F = feminine; GEN = genitive; INF = infinitive; IPFV = imperfective; M = masculine; MEDP = mediopassive; N = neuter; NGEN = non-genitive; NOM = nominative; PFV = perfective; PL = plural; PST = past; PTCP = participle; SG = singular.

[8]https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf

(3) **Nominal template**
stem:{gender}.{number}.{case}
ángelos 'angel:M.SG.NOM'
eγó '1.SG:NOM'

Nominal categories are added to the verbal categories of participial forms.

(4) **Verbal template**
stem:{voice}.{mood}.{aspect}.{tense}.{finiteness}
érθan 'come:PFV.PST.3.PL'
efovéθen 'fear:MEDP.PFV.PST.3.SG'
éxo γraménon 'have:1.SG write:PTCP:N.SG.ACC'

Additional transcription conventions were introduced to account for phenomena characteristic of spoken communication, as illustrated in Table 4.

xxx	unidentified words (mb/ge layer)
HESIT	hesitation
((laughs))	laughing

Table 4: Miscellaneous

6. Online Resources

The resources of this corpus contain three subcollections:

- **Subcollection TXT:** 24 speakers, 338 sound files (duration: 6h 7m), 35,843 words (Kotanidi et al., 2019).
- **Subcollection VA1:** 4 speakers, 30 sound files (duration: 38m), 3,830 words (Berikashvili, 2019).
- **Subcollection VA2:** 2 speakers, 5 sound files (duration: 34m), 3,400 words (Berikashvili and Skopeteas, 2019).

Audio files and transcriptions are available for download from the TLA archive (*The Language Archive*[9]) at the Max Planck Institute for Psycholinguistics in Nijmegen, a repository dedicated to data collections from endangered languages.

The corpus website[10] contains an overview of the resources, including the instructions used in data collection and the conventions for orthographic and morphological transcriptions. Users can navigate the resource and the audio files through a server installation of ANNIS (see Fig. 3), which is linked to the corpus website (Krause and Zeldes, 2016).

[9]https://archive.mpi.nl/tla/

[10]https://spw.uni-goettingen.de/projects/xtyp/PNT.html

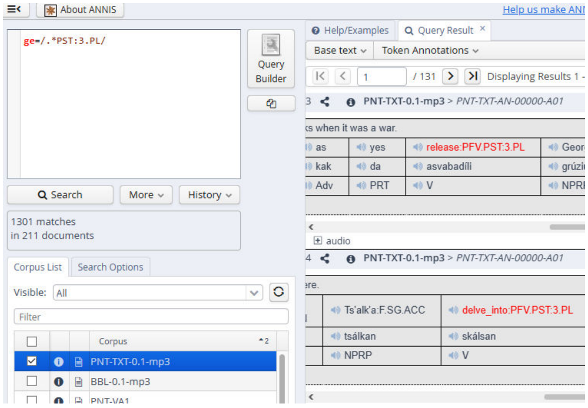

Figure 3: ANNIS interface for corpus queries

ANNIS provides the possibility to formulate queries over multi-layer annotations using its query language (AQL, ANNIS Query Language), which allows users to search for annotation attributes and relations using regular expressions. The illustrative query in Fig. 3 retrieves all tokens in the corpus that contain the string `PST:3.PL`, preceded by an arbitrary number (*) of arbitrary characters (.). Further example queries in AQL are provided on the corpus website.

7. Illustrating Variation in the Corpus Data

Pontic Greek people have always lived in multilingual settings, and the corpus reflects contact with several languages: (a) Turkish, including historical influence and later contact with Turkic varieties (e.g. Urum and Azerbaijani), especially in mixed regions and most strongly in texts from Dmanisi; (b) Standard Modern Greek, maintained through transnational family ties – particularly visible in Adjara; (c) Russian, with widespread bilingualism across regions; and (d) Georgian, which has become particularly influential following urban migration, especially to the capital.

The linguistic variation captured in the corpus is illustrated in the following examples. Figure 4 illustrates a text elicited with the CULTURE prompt (see §4.2). The excerpt, taken from a Pontic song, exhibits authentic Pontic lexical material and grammatical features, such as a subject in 'accusative' form: *xamelétan* 'mill:M.SG.ACC'.

The effects of language contact are also evident in lexical embeddings. Figure 5, for instance, demonstrates the embedding of a Russian noun in a Greek polydefinite. Currently, the language of origin (Russian, Turkish, Georgian) is only indicated in the Toolbox lexicon. In a latter version of our resource, this information can be exported to the ELAN files to be used in queries.

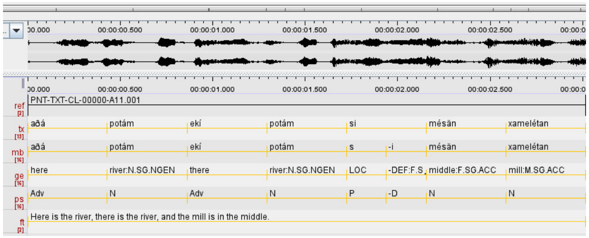

Figure 4: Pontic song (PNT-TXT-CL-00000-A11)

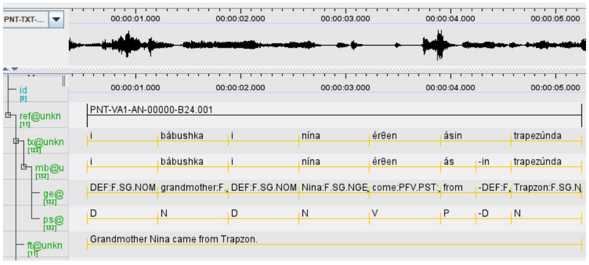

Figure 5: Russian embeddings in nominal structures (PNT-VA1-AN-00000-B24)

Figure 6 illustrates a typical strategy for integrating Russian verbs such as *prinimat'* 'take:IPFV.INF', namely through light-verb constructions (matrix verb: *epínen* 'make:IPFV.PST.3.SG'). The integration of verbs through light-verb constructions is cross-linguistically widespread (Wohlgemuth, 2009) and constitutes one of the main strategies for accommodating Russian loan verbs in Pontic Greek of Georgia (see Berikashvili, 2019 for a detailed discussion).

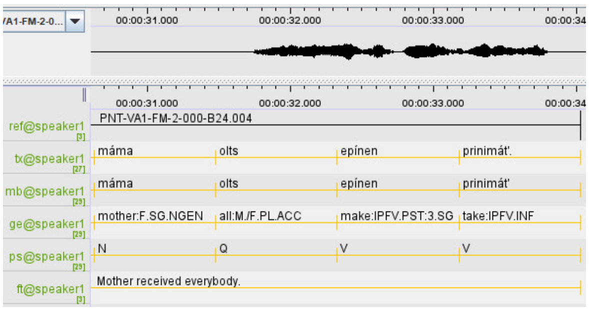

Figure 6: Russian embeddings in verbal structures (PNT-VA1-FM-2-000-B24)

The data in this corpus allows for quantitative assessments of effects of contact, such as the assignment of loanwords to inflectional classes or gender (see Berikashvili, 2022). Overall, the treatment of embeddings—both lexical and limited structural—places the variety represented in this corpus in the "more intense contact" category of the borrowing scale proposed by Thomason and Kaufman (1988).

8. Concluding Remarks

This article has presented an open-access, morphologically annotated multimedia corpus of Pontic Greek as spoken in the Caucasus, containing 373 recordings across three sociolinguistic stages that are representative for the settings in which the dialect is currently used. The resource fills a gap in the documentation of Pontic Greek outside Anatolia and contributes to our knowledge of an endangered dialect whose speaker base has drastically declined over recent decades. The corpus design – eliciting comparable narratives across speakers – facilitates systematic comparison between speakers and enables comparison between three important stages: homeland, internal migration, and external migration contexts.

The data highlights the complex multilingual ecology of the Caucasus, revealing patterns of contact with Russian, Georgian, Turkish, and Standard Modern Greek, and the resulting hybrid linguistic profiles. Beyond its relevance for Greek dialectology, the corpus provides valuable material for research on language contact, possible phenomena of koineization, and dialect attrition. We hope that this resource will serve both as a foundation for future empirical studies and as a contribution to the broader effort of safeguarding the linguistic heritage of Pontic communities in the Caucasus.

9. Acknowledgements

The corpus recordings, as well as the original transcriptions and translations of the main subcollection, were produced by Evgenia Kotanidi (Tbilisi). Svetlana Berikashvili revised the transcriptions and translations, added morphological analyses and glosses, and created two additional subcollections that form part of this corpus. Stefanie Böhm and Johanna Lorenz supervised the data collection and transcription process. The creation of this corpus was funded by the Volkswagen Foundation. We are grateful to two anonymous reviewers for their helpful comments and suggestions on an earlier version of this article. Responsibility for the present article rests with the authors.

10. Bibliographical References

Shukia Apridonidze and Levan Chkhaidze. 2002. From Georgian and into Georgian: Transliteration of Georgian alphabet. Linguistics Institute of Academy of Science of Georgia.

Svetlana Berikashvili. 2017. *Morphological Aspects of Pontic Greek Spoken in Georgia*. LINCOM, Munich.

Svetlana Berikashvili. 2019. Loan verb adaptation in Pontic Greek spoken in Georgia. In *Language contact in the Balkans and Asia Minor*, pages 262–279, Thessaloniki. Institute of Modern Greek Studies.

Svetlana Berikashvili. 2022. Contact-induced change in the domain of grammatical gender in Pontic Greek spoken in Georgia. *Languages*, 7:79.

Stergios Chatzikiriakidis, Chatrine Qwaider, Ilias Kolokousis, Christina Koula, Dimitris Papadakis, and Efthymia Sakellariou. 2025. Grdd: A dataset for Greek dialectal nlp. Version v5.

Georges Drettas. 1997. *Aspects pontiques*. Association de Recherches Pluridisciplinaires, Paris.

Fatima Eloeva. 1997. *Pontijskij dialekt (na materiale grecheskih bespis'mennyh govorov Gruzii i Krasnodarskogo kraja) [Pontic dialect (based on data of oral Greek varieities spoken in Georgia and Krasnodar Krai]*. St. Petersburg University, St. Petersburg.

Katherine Hodgson. 2010. Morphological evidence for the origins of a Pontic Greek community in armenia. *Studies in Greek Linguistics*, 30:237–248.

Concha Maria Höfler. 2020. *Boundaries and Belonging in the Greek Community in Georgia*. Nomos, Baden-Baden.

Nikitas N. Karanikolas, Eleni Galiotou, Dimitris Papazachariou, Konstantinos Athanasakos, George Koronakis, and Angela Ralli. 2015. Towards a computational processing of oral dialectal data. In *PCI '15: 19th Panhellenic Conference on Informatics*, pages 337–341.

Simon Kaukhchishvili. 1942. Berdznebis dasakhlebis ist'oria sakartveloshi [the history of Greek settlement in Georgia]. *Works of the A. Tsulukidze Kutaisi State Pedagogical Institute*, IV:219–239.

Simon Kaukhchishvili. 1946. Berdznebis dasakhlebis ist'oria sakartveloshi [the history of Greek settlement in Georgia]. *Works of the A. Tsulukidze Kutaisi State Pedagogical Institute*, VI:125–153.

Thomas Krause and Amir Zeldes. 2016. Annis3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118—139.

Ekkehard König, Dik Bakker, Östen Dahl, Martin Haspelmath, Maria Koptjevskaja-Tamm, Christiam Lehman, and Anna Siwierska. 1993. *Eurotyp Guidelines*. European Science Foundation Programme in Language Typology.

Evstathios Lazarev. 1950. *Calkinskij dialekt grecheskogo jazyka (foneticheskij i morfologicheskij analiz) [Tsalka dialect of Greek (phonetic and morphological analysis)]*. Ph.D. thesis, University of Tbilisi.

Nika Loladze. 2021. *Migratory Movements of Georgia's Greek Community. The Impact of Current Socio-economic transformations*. Peter Lang, Berlin.

Peter Mackridge. 1987. Greek-speaking moslems of North-East Turkey: Prolegomena to a study of the Ophitic sub-dialect of Pontic. *Byzantine and Modern Greek Studies*, 11(1):115–137.

Dimitra Melissaropoulou, Stavros Bompolas, and Charalampos Tsimpouris. 2022. Digital cartography in the service of preservation of cultural linguistic heritage: Implementing the electronic dialectal atlas of Cappadocian Greek. *Scientific Culture*, 8(2):135–146.

Avtandil Mikaberidze and M. Shakhpazidi. 2000. Sakartveloshi mtskhovreb berdzenta dialekt'isatvis [about the dialect of Greeks living in Georgia]. In *Berdznebi sakartveloshi [Greeks in Georgia]*, pages 129–178. Logos, Tbilisi.

Dimosthenis Oikonomidis. 1958. *Grammatikī tīs ellinikīs dialektou tou Pontou [Grammar of the Greek Pontic dialect]*. Academy of Athens, Athens.

Anthimos A. Papadopoulos. 1955. *Istorikī grammatikī tīs Pontikīs dialektou [Historical grammar of the Pontic dialect]*. Committee for Pontic Studies, Athens.

Anthi Revithiadou and Vassilios Spyropoulos. 2012. *Ophitikī. Ptykhes tīs grammatikīs domīs mias pontiakīs dialektou [Ofitika. Aspects of grammatical structure of the Pontic dialect]*. Publishing house of brothers Kyriakidi, Thessaloniki.

Laurentia Schreiber. 2021. Pontic Greek (Romeyka). In G. Haig, D. Stilo, M. C. Doğan, and N. N. Schiborr, editors, *WOWA - Word Order in Western Asia. A spoken-language-based corpus for investigating areal effects in word order variation*. University of Bamberg, Bamberg.

Eleni Sideri. 2006. *The Greeks of the former Soviet Republic of Georgia. Memories and Practices of Diaspora*. Ph.D. thesis, University of London.

SIL. 2015. Field linguist's toolbox. [Computer software].

Ioanna Sitaridou. 2013. Greek-speaking enclaves in Pontus today: The documentation and revitalization of Romeyka. In M. C. Jones and S. Ogilive, editors, *Keeping languages alive, documentation, pedagogy, and revitalization*, pages 98–112. Cambridge University Press, Cambridge.

Han Sloetjes and Pitter Wittenburg. 2008. Annotation by category: Elan and iso dcr. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 816–820, Marakech, Morocco. European Language Resource Association (ELRA).

Sarah Grey Thomason and Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. University of California Press, Berkley, Los Angeles, London.

Dimitris Tombaidis. 1988. *Ī Pontiakī dialektos. Dialektika charaktīristika, katataxī idiōmatōn, dialektika keimena [Pontic dialect. Dialectic characteristics, classification and texts]*. Committee for Pontic Studies, Athens.

Jan Wohlgemuth. 2009. *A Typology of Verbal Borrowings*. Mouton de Gruyter, Berlin.

Artemis Xanthopoulou-Kyriakou. 1997. Metoikesies ton ellinōn tou pontou pros tis xōres tou kavkasou (1829-arxes 20ou aiōna) [migration of Greeks from Pontos to Caucasian countries (1829—beginnings of the 20th century)]. In I. Khasiotis, editor, *Oi Ellīnes tīs Rōsias kai tīs Sovietikīs Enōsīs. Metoikesies kai ektopismoi. Organōsī kai ideologia. [Greeks from Russia and Soviet Union. Migration and expatriation. Organization and Ideology]*, pages 85–127. University Studio Press, Thessaloniki.

11. Language Resource References

Berikashvili, Svetlana. 2019. *Pontic Data Collection 2*. The Language Archive, Corpus resource, 3.0. PID https://hdl.handle.net/1839/00-0000-0000-0021-4DA4-3.

Berikashvili, Svetlana and Skopeteas, Stavros. 2019. *Pontic Data Collection 3*. The Language Archive, Corpus resource, 3.0. PID https://hdl.handle.net/1839/00-0000-0000-0021-4DA4-3.

Kotanidi, Evgenia and Berikashvili, Svetlana and Böhm, Stefanie and Lorenz, Johanna and Skopeteas, Stavros. 2019. *Pontic Data Collection*. The Language Archive, Corpus resource, 3.0. PID https://hdl.handle.net/1839/00-0000-0000-0021-4DA4-3.