Pontic Greek in the Caucasus: an online corpus

Svetlana Berikashvili, Stavros Skopeteas

Ilia State University (1), University of Göttingen (2)
Kakutsa Cholokashvili Ave 3/5, Tbilisi 0162, Kate-Hamburger-Weg 3, Göttingen 37073
svetlana.berikashvili@iliauni.edu.ge, stavros.skopeteas@uni-goettingen.de

Objectives

The main objective is to present a multi-media corpus of Pontic Greek as spoken in the Caucasus (Georgia). The corpus covers three major stages reflecting different sociolinguistic settings:

- **Stage A:** HOMELAND: rural communities in the original settlements in Georgia;
- **Stage B:** INTERNAL MIGRATION to urban centers within Georgia;
- **Stage C:** EXTERNAL MIGRATION to Greece.

Dataset (open-access resource):
- 373 audio recordings
- total duration 7h 26m
- word count: 43,073 words

Introduction

The corpus is designed to capture variation in Pontic Greek, as spoken by the Pontic speakers of Georgia. Pontic Greek in the Caucasus remains underrepresented in research on Greek dialects, and in light of migration processes, has become severely endangered. Its systematic documentation in online resources is therefore an urgent priority.

Pontic speakers migrated to the Caucasus from various areas of Anatolia (see Figure 1) largely beginning in the 19th century. They are dialectally heterogeneous (see places of origin), and continue speaking Pontic in a multilingual environment (major languages: Russian, Georgian).

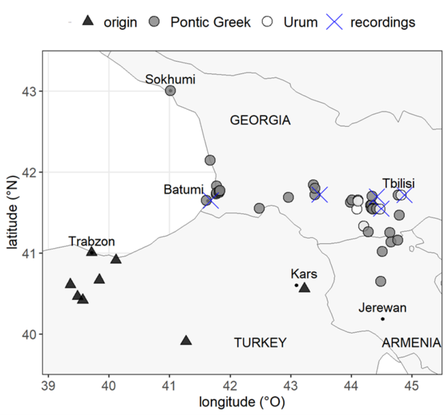

Figure 1: Greek populations in the Caucasus

This corpus offers a unique opportunity to investigate contact between Greek and Indo-European (Russian) as well as Non-Indo-European languages (Georgian, Turkish).

Beyond the linguistic relevance, the materials also provide insights into cultural practices. The resource is available both as an online corpus:
https://spw.uni-goettingen.de/projects/xtyp/PNT.html
and to download (TLA archive):
https://hdl.handle.net/1839/00-0000-0000-0021-4DA4-3

Data Collection

The data have been collected in Georgia and Greece. The data collection included the following stages:

- identifying speakers in three sociolinguistic settings;
- instructing speakers by native speaker and obtaining written informed consent;
- conducting interviews in non-laboratory environments.

BETWEEN-SPEAKERS DESIGN: all speakers produced narratives on the same topics (identical instructions): 'Ancestors', 'Family', 'Village', 'Culture', 'People', 'Marriage', 'Feast' and 'Language'.

METADATA: information on the date and location of recordings, as well as speaker-related variables (See Figure 2 for self-reported frequency of language use).

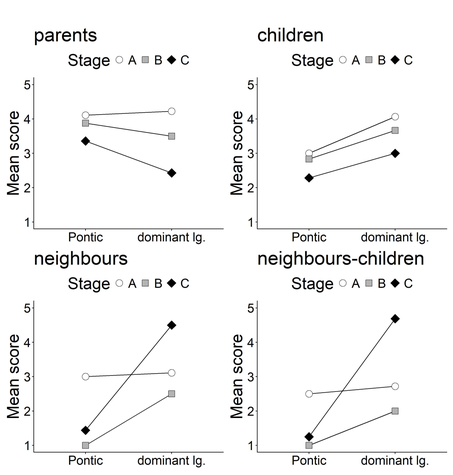

Figure 2: Aggregated self-estimation scores of the language use with different addressees

Annotations

Data processing included the following stages:

- transcription;
- morphological annotation and glossing in Toolbox;
- exporting of Toolbox annotations to ELAN, where sentence boundaries were time-aligned with the audio files.

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. The corpus has a multi-layer annotation structure; see tiers in Table 1.

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
ge_a	mb	*gloss-English_associated:* original word-by-word translation
ps	mb	*part of speech:* part of speech classification
ft	ft	*free translation:* sentence-by-sentence translation
nt	ref	*notes:* free comments

Table 1: Annotation tiers

TRANSCRIPTIONS followed orthographic conventions that are generally close to a broad phonological transcription.

MORPHOLOGICAL ANNOTATIONS are based on *The Leipzig Glossing Rules* with additions of abbreviations from *Eurotyp*.

Online Resources

The resources of this corpus contain three subcollections:

- **Subcollection TXT:** 24 speakers, 338 sound files (duration: 6h 7m), 35,843 words [1].
- **Subcollection VA1:** 4 speakers, 30 sound files (duration: 38m), 3,830 words [2].
- **Subcollection VA2:** 2 speakers, 5 sound files (duration: 34m), 3,400 words [3].

The CORPUS WEBSITE provides 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 Figure 3).

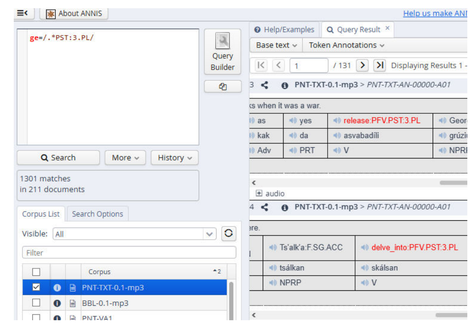

Figure 3: ANNIS interface for corpus queries

Concluding Remarks

The resource fills a gap in the documentation of Pontic Greek outside Anatolia and may serve:

- 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.

Pontic Greek Corpus

The Pontic Greek dataset comprises a morphologically annotated corpus of narrative texts and additional interviews. The corpus website allows users to search for annotation attributes and relations using AQL (ANNIS Query Language) that provides the possibility to formulate queries over multi-layer annotations.

References

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

[2] Svetlana Berikashvili.
Pontic Data Collection 2, The Language Archive, 2019.
https://hdl.handle.net/1839/00-0000-0000-0021-4DA4-3.

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