Corpus of Pontic Greek Dialect as spoken in Georgia and Digital Principles of Linguistic Research

Svetlana Berikashvili

Tbilisi State University

19-20 September, 2016

Outline

- Introductory part;
- Presentation of the collected data;
- Presentation of the digital software used for the annotation;
- TLA archive and its use in linguistic research;
- Results of the implemented work.

Introduction

Greek community in Georgia

- Urum speakers;
- Pontic Greek speakers;

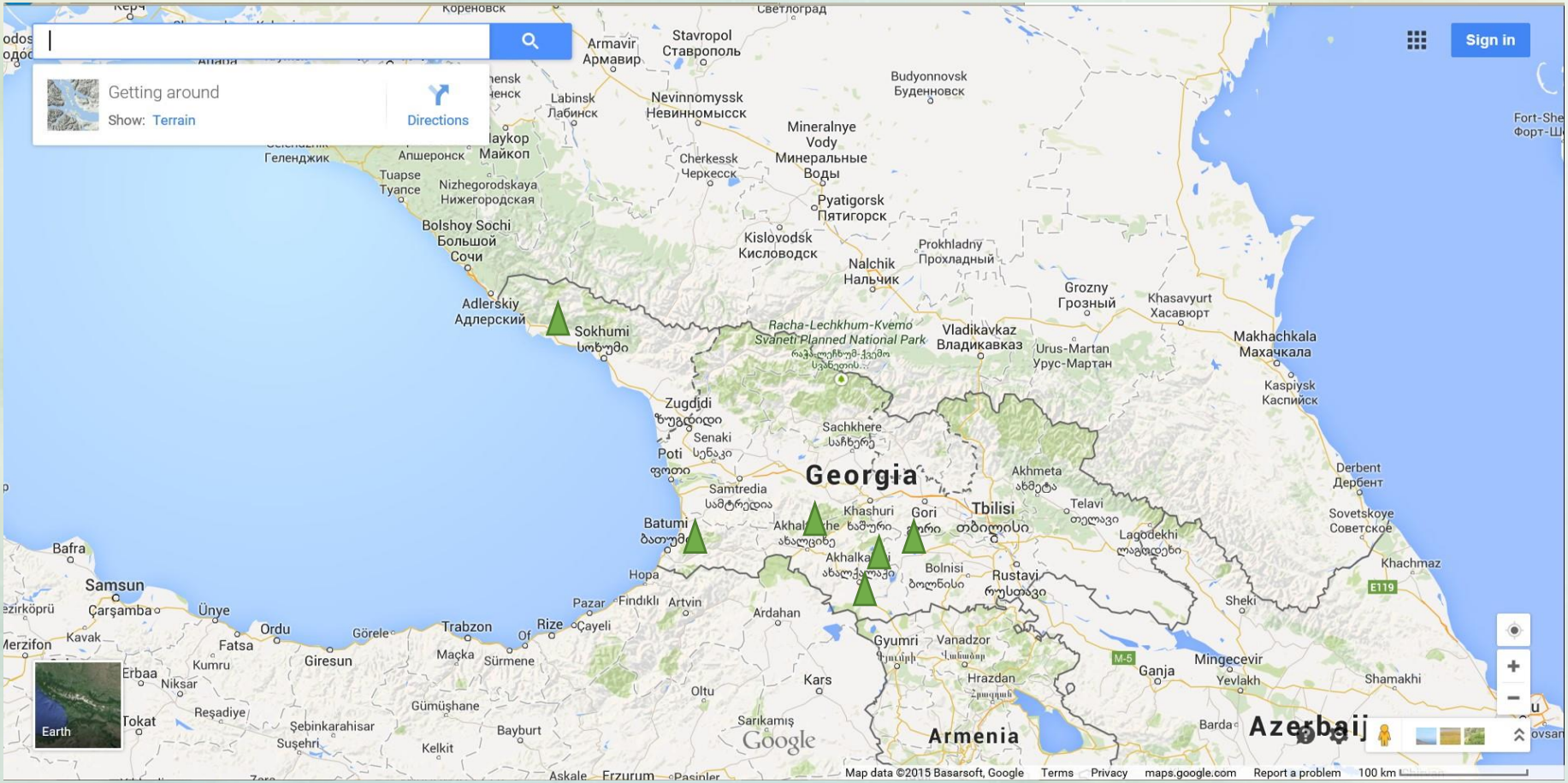
Three different stages

- Stage A: Homeland – original settlements of Pontic Greek speakers in Georgia

- Stage B: Internal migration – from original settlements to the cities, generally to capital of Georgia, Tbilisi

- Stage C: Emigration – to the countries of European Union, mostly in Greece

Introduction



Collected data

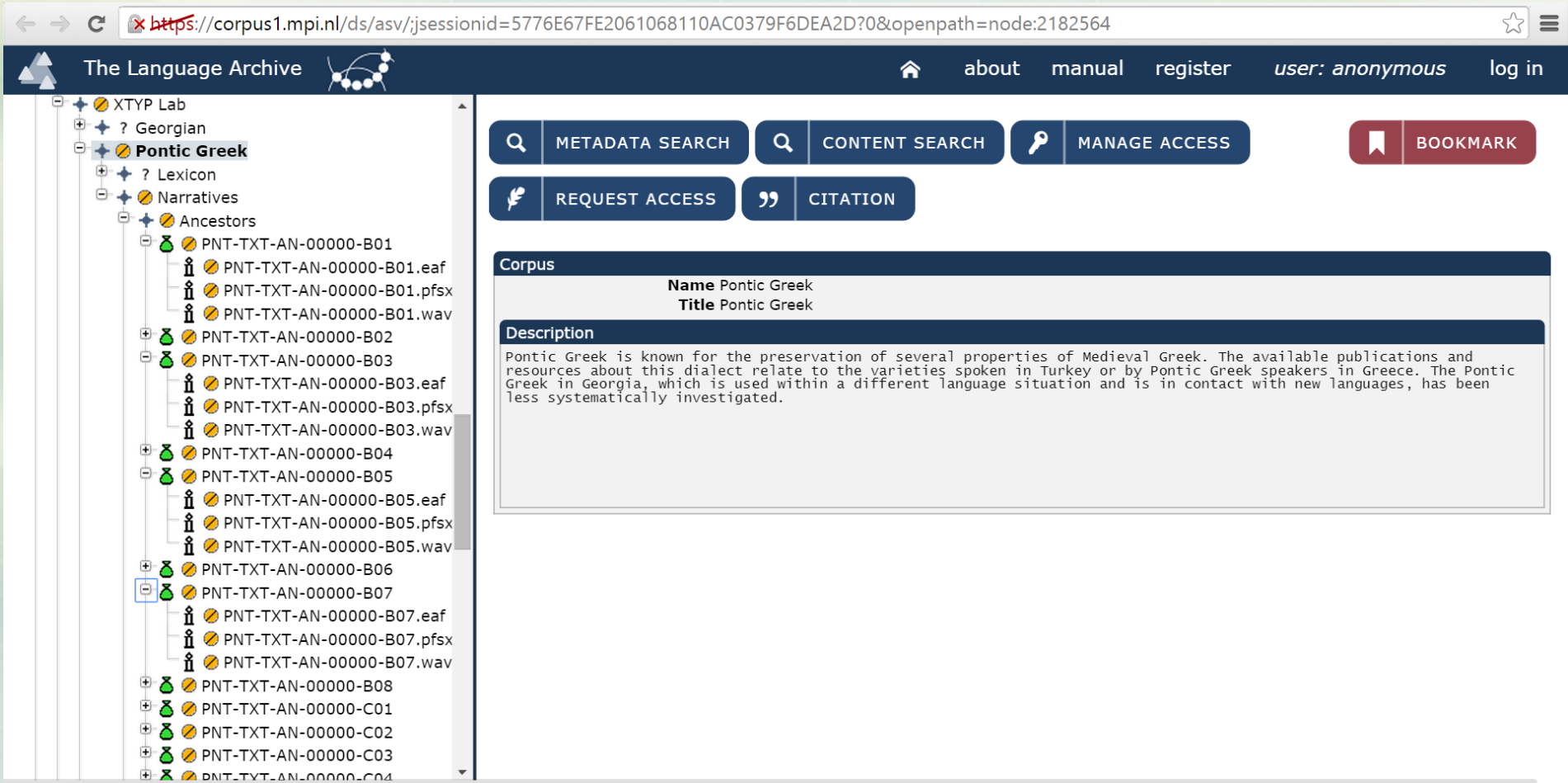
Data

- Multimedia corpus, uploaded to the TLA archive;
- Quantity of data: 435 media and ELAN files, the average word count per speaker is 936 words, 57 native-speaking informants, in total the corpus contains 53 295 words;
- Narratives on the different (8) topics, namely Ancestors, Family, Village, Culture, People, Marriage, Feast, Language.

Data recoded by Stavros Skopeteas, Evgenia Kotanidi and Svetlana Berikashvili

Data glossed by Svetlana Berikashvili

TLA archive


https://corpus1.mpi.nl/ds/asv/;jsessionid=5776E67FE2061068110AC0379F6DEA2D?0&openpath=node:2182564
The Language Archive
about
manual
register
user: anonymous
log in
XTYP Lab
? Georgian
Pontic Greek
? Lexicon
Narratives
Ancestors
PNT-TXT-AN-00000-B01
PNT-TXT-AN-00000-B01.eaf
PNT-TXT-AN-00000-B01.pfsx
PNT-TXT-AN-00000-B01.wav
PNT-TXT-AN-00000-B02
PNT-TXT-AN-00000-B03
PNT-TXT-AN-00000-B03.eaf
PNT-TXT-AN-00000-B03.pfsx
PNT-TXT-AN-00000-B03.wav
PNT-TXT-AN-00000-B04
PNT-TXT-AN-00000-B05
PNT-TXT-AN-00000-B05.eaf
PNT-TXT-AN-00000-B05.pfsx
PNT-TXT-AN-00000-B05.wav
PNT-TXT-AN-00000-B06
PNT-TXT-AN-00000-B07
PNT-TXT-AN-00000-B07.eaf
PNT-TXT-AN-00000-B07.pfsx
PNT-TXT-AN-00000-B07.wav
PNT-TXT-AN-00000-B08
PNT-TXT-AN-00000-C01
PNT-TXT-AN-00000-C02
PNT-TXT-AN-00000-C03
METADATA SEARCH
CONTENT SEARCH
MANAGE ACCESS
BOOKMARK
REQUEST ACCESS
CITATION
Corpus
Name Pontic Greek
Title Pontic Greek
Description
Pontic Greek is known for the preservation of several properties of Medieval Greek. The available publications and resources about this dialect relate to the varieties spoken in Turkey or by Pontic Greek speakers in Greece. The Pontic Greek in Georgia, which is used within a different language situation and is in contact with new languages, has been less systematically investigated.

Annotation of the data

- Transcription of the files is base on orphographic convention as close as possible to a broad phonological transcription, namely Pontic Greek uses convention based on the IPA-phonemic transcription in Drettas (1997)

Vowels

Consonants

Consonant inventory (IPA values in brackets; orthography in italics)

	vowel	example			postalv.	palatal	velar
plosive		IPA	orthography	meaning		[c] k	[k] k
	[a]	[calaʧi]	kalachí	‘conversation’		[ɟ] g	[g] g
fricative					[ʃ] sh		[x] x
	[æ] (ą)	[ðævolos]	ðä´volos	‘devil’			
	[e]	[eksero]	ekséro	‘know’	[ʒ] zh	[j] j	[ɣ] ɣ
affricate	[o]	[omati]	omáti	‘eye’	[ʧ] ch		
	[i]	[imera]	iméra	‘day’	[ʒ] dzh		
nasal	[u]	[ɣurban]	ɣurbán	‘sacrifice’			[ŋ] n
tap			[ɾ] r				
lateral			[l] l				[ɬ] l
approximant						[j] j	

Annotation of the data

The morphological transcription

- Leipzig Glossing Rules
https://www.eva.mpg.de/lingua/resources/glossing-rules.php

- Eurotype 1983
https://www.eva.mpg.de/lingua/tools-at-lingboard/pdf/eurotyp_guidelines/Eurotyp_GL_chapter3_Glossing.pdf

- ISIC 2007 (Information Structure in Cross-Linguistic Corpora)
https://pub.uni-bielefeld.de/publication/2144080

Glossing Principles for PG

a) Nominals (Adjectives, Substantives, Pronouns, adjectival Participles) - (Gender, Number, Case)

Gender: M, N, F
Number: SG, PL
Case: NOM, GEN, ACC, VOC; NGEN

e.g. *ángelos*
angel:M.SG.NOM,

eɣó
1.SG:NOM

Glossing Principles for PG

b) Verbs - (Voice, Mood, Aspect, Tense, Finiteness, Person/Number)
As a general rule: we do not indicate unmarked categories: active, indicative, present, finite verb

érθan	come:PFV.PST.3.PL
efovéθen	fear:MEDP.PFV.PST.3.SG
píson	do:IMP.PFV.2.SG
éxo ɣraménon	have:1.SG write.PTCP:N.SG.ACC

Voice: MEDP (=mediopassive)
Mood: IMP (=imperative), SUBJ (=subjunctive), OPT (=optative)
Aspect: IPFV (=imperfective), PFV (=perfective)
Tense: PST (=past), FUT (=future), PRF (=perfect, synthetic)
Finiteness: INF (=infinitive), PTCP (=participle)
Person/Number: 1, 2, 3, SG, PL

Glossing Principles for PG

c) Part of Speech is indicated as follows: N=noun, V=verb, A=adjective, Adv=adverb, P=adposition, Q=quantifier, AQ=ordinal nominals, C=conjunction, PN=pronoun, PRT=particle, X=unclear.

d) Diminutive nouns with endings: *-aki, -ópon* etc. are glossed as follows, e.g. *pulákia* bird.DIM:N.PL.NGEN

e) As Future is denoted by using the particle *θa* and dependent / independent form of the verb (dependent that of the Perfective aspect), the glossing is as follows, e.g.

θa	*páte*
FUT	go:DEP:2.PL

The same is with the Optative mood, e.g.

as	*émna*
OPT	be:IPFV.PST.1.SG

Glossing Principles for PG

f) The same form is used to indicate Subjunctive and Indicative form of the verb, simply Subjunctive has particle *na*, which is glossed separately as particle. In this case we do not indicate SUBJ. and IND., e.g. *na ékserna* – to know:IPFV.PST:1.SG

g) In case of the possessive pronouns, *temón, temá, teméteron, temétera* which denote the number of the head noun as well, this is also indicated in gloss, e.g. *temón i ɣlósa* - POSS.1.SG:N.SG, *temá taðélfia* - POSS.1.SG:N.PL. The same is for the pronouns *teméteron* and *temétera.* So, the glosses are as follows:

t=emón
DEF=POSS.1.SG:SG

t=emá
DEF=POSS.1.SG:PL

t=eméteron
DEF=POSS.1.PL:SG

t=emétera
DEF=POSS.1.PL:N.PL.NGEN

t=eméteri
DEF=POSS.1.PL:M/F.PL.NOM

t=emetér
DEF=POSS.1.PL:M/F.PL.NOM

Glossing Principles for PG

h) *t-emón* : this means that the article is a prefix.
t=emón : this means that article and pronoun form a phonological word. i.e., the article is a preclitic
In our case the second option is correct, but the symbol = cannot be used in toolbox, that's why in toolbox is used – for reasons of convenience.

i) In forms like *pápasim, mámasim* the letter *i* appears after a word that ends in a consonant, mainly *s*, it is added before clitics *mu, su* and is reanalysed as part of the clitic. The gloss is as follows:

pápas=im	*mámas=im*
father:M.SG.NOM=CL.1.SG:GEN	mother:F.SG.GEN= CL.1.SG:GEN

j) Comperative and Superlative forms are indicated, e.g. *kirióteron* - main.CMPR:N.SG.NGEN

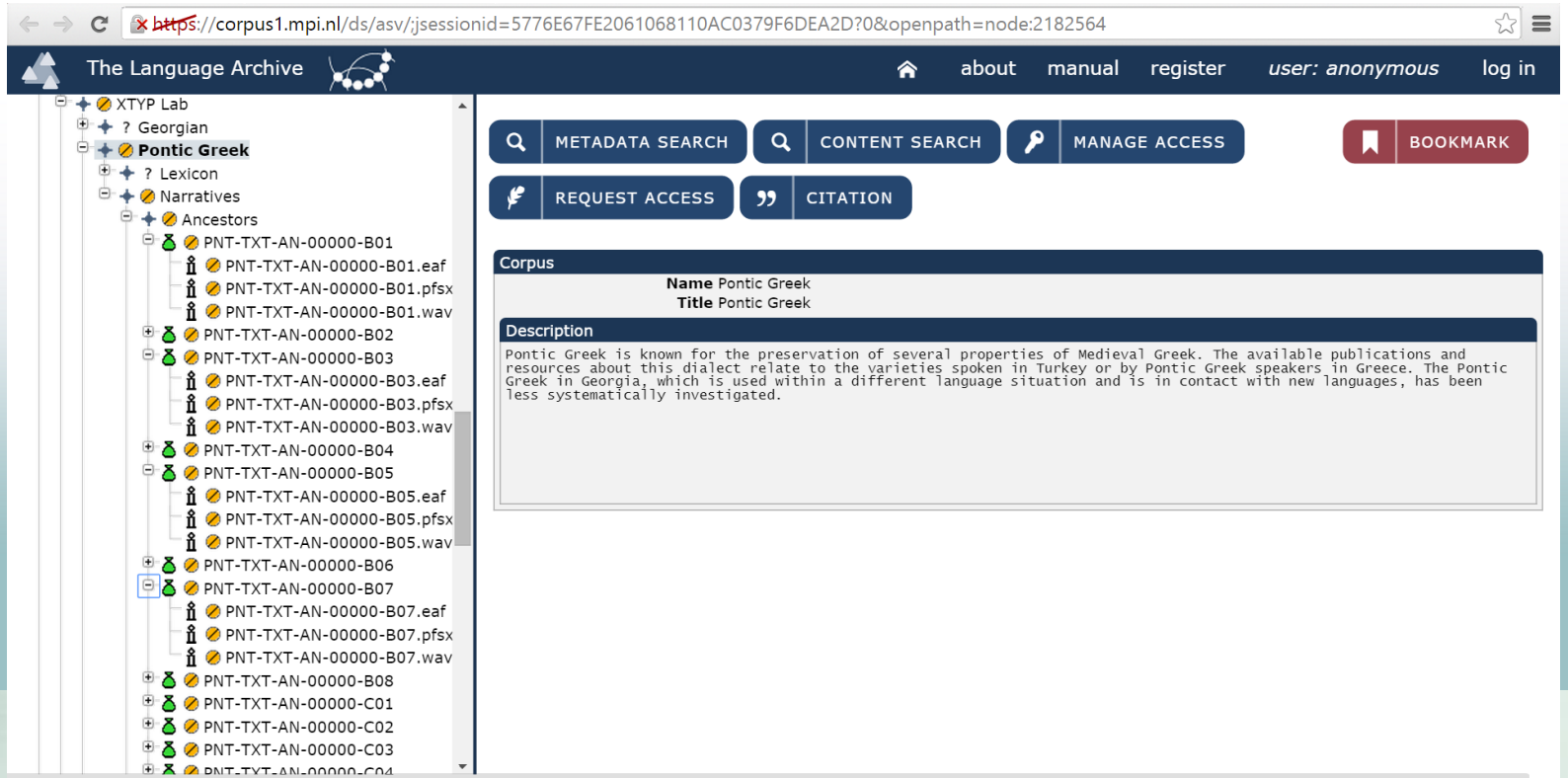
Software used for the annotation

- Toolbox
http://www-01.sil.org/computing/toolbox/

- ELAN
https://tla.mpi.nl/tools/tla-tools/elan/

- LAMUS
https://tla.mpi.nl/wp-content/uploads/2012/01/lamus2_workspace.png

TLA archive



https://tla.mpi.nl/resources/data-archive/

Results and further use

The result of the implemented work is a big multi-media annotated corpus of Pontic Greek variety as spoken in Georgia, so called Romeyka dialect, which includes narratives of approximately 53 295 words, produced by 57 native-speaking informants.

The priority of this data collection is that it is archived according to current sustainability standards of linguistic resources thus it is available to the research community, and will remain accessible even after completion of the current project.

Through integration into the TLA archive the data can be used for the research and teaching purposes, thus linking research and teaching, which is very important in teaching nowadays. The use of the data can promote research-based teaching, and help to promote initiatives in this direction. It is already used by the students of the Institute of Classical, Byzantine and Modern Greek Studies, TSU in their linguistic and sociolinguistic researches.

References: Data base resources

BERIKASHVILI, SVETLANA 2016. *Interviews in Pontic Greek* (Data collected, transcribed, and glossed by Svetlana Berikashvili). Bielefeld: Bielefeld University (corpus resource: TLA, Donated Corpora, XTYP Lab).

KOTANIDI, EVGENIA, SVETLANA BERIKASHVILI, STEFANIE BÖHM & JOHANNA LORENZ, STAVROS SKOPETEAS. 2016. Pontic data collection, Version 2.0 (data collected, transcribed, and translated by Evgenia Kotanidi; data glossed by Svetlana Berikashvili; supervised by Stefanie Böhm and Johanna Lorenz; corpus design by Stavros Skopeteas). Bielefeld: Bielefeld University (corpus resource: TLA, Donated Corpora, XTYP Lab).

SKOPETEAS, STAVROS & SVETLANA BERIKASHVILI 2016. *Interviews in Pontic Greek* (Data collection by Stavros Skopeteas, 2005; transcribed, translated and glossed by Svetlana Berikashvili, 2016). Bielefeld: Bielefeld University (corpus resource: TLA, Donated Corpora, XTYP Lab).