Pontic project: Theoretical problems and modern technologies

Maxim Kisilier

Saint Petersburg University, Institute for Linguistic Studies RAS

Οι σύγχρονες ψηφιακές τεχνολογίες ανοίγουν νέους ορίζοντες στη φιλολογία και γλωσσολογία. Παρ' όλο που οι νεοελληνιστές δεν είναι ειδήμονες στη χρήση των νέων τεχνολογιών, έχουν δημιουργηθεί τουλάχιστον πέντε διαφορετικά σώματα κειμένων της ελληνικής γλώσσας, ενώ, τελευταία, επιστήμονες που ασχολούνται με τις νεοελληνικές διαλέκτους ψάχνουν τρόπους να εκμεταλλευτούν τα αποτελέσματα της τεχνολογικής ανάπτυξης. Το άρθρο αυτό αφορά την ποντιακή διάλεκτο στη Ρωσία και στις χώρες της πρώην Σοβιετικής Ένωσης, όπου οι ποντιόφωνοι είναι περισσότεροι από 20.000, χωρίς η γλώσσα τους να έχει μελετηθεί αρκετά. Πιο συγκεκριμένα, στο άρθρο περιγράφεται η βάση δεδομένων που δημιουργείται με σκοπό την ανάλυση διαφόρων παραλλαγών της Ποντιακής, καθώς και τη σύγκρισή της με τις άλλες νεοελληνικές διαλέκτους.

Keywords: Greeks of the USSR, Pontic Greek, corpus linguistics, dialectal database

1. Introduction: Modern technologies in MG studies

Technological achievements of the 20th century have radically changed methods and approaches in many scientific areas. Even those who are far from modern physics and astronomy have heard about the NASA Hubble Space Telescope or the Large Hadron Collider. By contrast, the humanities (especially philology and linguistics) may appear insensitive to the progress if the conversion from pen to MS Office is not taken into account.

This impression is totally false. Nearly for all branches of linguistics and philology special software is being constantly developed, like Praat for phonetics and phonology, TLex for lexicography, FieldWorks for morphology, LingTree for syntax, etc. However, the creation of language corpora has become the most important breakthrough. Although Modern Greek studies are not in the forefront of technological innovation, there are at least 5 corpora of Modern Greek[1]:

---

[1] It is also worth mentioning the Thesaurus Linguae Graecae (TLG), the corpus of Ancient Greek, which is one of the most powerful language corpora.

(i) Corpus of Greek Texts (CGT) — 30 million tokens (cf. Goutsos 2010)

(ii) Corpus of Spoken Greek of the Institute of Modern Greek Studies (CSG) — over 2 million tokens (cf. Pavlidou 2016)

(iii) Hellenic National Corpus (HNC) — 47 million tokens

(iv) the Greek Web Corpus (elTenTen) — 1.6 billion tokens

(v) Corpus of Modern Greek (CMG) — 37.5 million tokens (cf. Arkhangelskiy & Kisilier 2018)

They are not similar both in the selection of texts and in technical options (for brief overview of their functionality see Table 1). Each corpus has its own advantages and drawbacks and none of them makes use of dialectal data.

	CGT	CSG	HNC	elTenTen	CMG
Free access	+	+	−	−	+
Lemma search	−	−	+	+	+
Grammar search	−	−	+/−	+/−	+
Synonyms	−	−	−	+	−
Concordance	−	−	+	+	−
Lemmas' English translation	−	−	−	−	+
Metadata	+/−	+	+	−	+
Gender studies friendly	−	+	−	−	−
Distance between constituents of the required phrase	−	−	+	−	+
Search in particular text/author	−	−	+	−	+
Grammatical analysis of a particular word	−	−	−	−	+

Table 1: Corpora of Modern Greek

Modern Greek dialectology has also shown recently some progress in implementation of digital technologies. The Laboratory of Modern Greek dialects of the University of Patras has developed a set of databases which can be transformed into dialectal corpora, atlases and dictionaries (for details see Ralli 2015). Unfortunately, most researchers cannot afford such kind of projects and look for other possible solutions, for example Praat scripts, etc. However, it is important to elaborate a unified approach which would result in multiple compatible databases of various Modern

Greek dialects. In this contribution a possible approach will be discussed using data from Pontic Greek in the USSR and Russia.

2. Pontic Greek in the USSR and Russia

Pontic Greeks are the most widespread and large Greek diaspora and the list of countries which host its representatives seems endless. For example, in Greece the number of Pontic Greek speakers is roughly estimated at 300.000 (Sitaridou & Kaltsa 2014: 3). Even in Turkey, there are still about 5000 speakers of Pontic (Schreiber & Sitaridou 2018: 2). However, these calculations may be underestimating. According to Hakan Özkan (2013: 2), during the Census of 1965, in the province of Trabzon 4535 persons claimed that Pontic is their mother language, but the same Census demonstrates that there were 5740 inhabitants living only in Pontic-speaking villages of Beşköy (and, evidently, most of them could speak Pontic). It means that many speakers did not indicate (or did not want to indicate) Pontic as their L1. Since (nearly) all speakers of Pontic are not monolingual, the situation described by Hakan Özkan may be relevant for other countries as well.

Despite the number of speakers, Pontic in general can be identified as endangered according to the classification by Moseley (2007: xi): the youngest generation who can speak Pontic fluently are young adults. And nevertheless, Pontic is more widely used for communication than most Modern Greek dialects. It happens because Pontic Greeks from different countries frequently communicate with each other and Pontic is the only language everyone understands. So, it is not quite correct to describe the use of Pontic today exclusively in terms of cultural and national identity (cf. Konstantinidou & Kyridis 2014: 633).

2.1. Historical overview

First waves of migration to the Russian Empire date back to the 1820s[2]. For example, Pontic village Anadol (not far from Mariupol) was founded in 1826 (Animitsa & Kisilier 2009: 31). Due to the Asia Minor Catastrophe, multiple refugees from

---

[2] In the 15th century and later, Trabizond Greeks were regularly relocated to the Crimea (Sokolov 1932: 313).

Pontos appeared in the Caucasus (cf. Der Matossian 2009), the Crimea, and in Southern Russia (find more details in Vergeti 1991). Initially, Pontic Greeks did not have literary tradition, but in the late 1920s, the Soviet Union launched the National policy on linguistic minorities. Thus, Pontic Greeks got their alphabet. It was based on Demotic Greek and it was created for all Greek communities in the USSR (Table 2, cf. Savvov 1931).

A α	E ε	K κ	O o	T τ
B β	Z ζ	Λ λ	Π π	Υ υ
Γ γ	Θ θ	M μ	P ρ	Φ φ
Δ δ	I ι	N ν	Σ ς	X χ

Table 2: Soviet Greek alphabet

This alphabet retained only one way to express /e/ (= <ε>), /i/ (= <ι>) and /o/ (= <o>). The grapheme <υ> was used for /u/, while digraphs <αι>, <ει>, <οι> and <ου> were no longer required. Instead of <αυ> and <ευ> it was supposed to write <αβ/αφ> and <εβ/εφ>. New orthography avoided geminates[3], <ζ> could be both /z/ and /ʒ/, <ςς> was /ʃ/ and <τςς> — /t͡ʃ/.

From the very beginning, the Soviet Pontic literature had to compete with the Soviet Demotic Greek literature which was vigorously supported by multiple Greek communists who came to the USSR. As a result, Demotic Greek became the language of Greek schools although it was incomprehensible to the speakers of the local Greek dialects[4]. However, during 1930s, a number of books in Pontic were published (cf. Kachalov 1934).

After 1938, literature in Greek was prohibited and some groups of Pontic Greeks were forcibly displaced to Uzbekistan, Kazakhstan and Siberia. Interestingly, a considerable part of Pontic Greeks became Soviet citizens only in 1970s[5]. Nowadays, Pontic communities exist in various regions of Russia including Moscow and Saint Petersburg, but the dialect is mostly spoken in Southern Russia and the Caucasus. According to Sam Topalidis (2018: 10), in 2010, there were 85.600 Pontic Greeks in Russia and about 23.000 of them could speak the dialect.

---

[3] In a way it was a realization (although unconscious, I believe) of orthographic ideas of Nikos Kazantzakis expressed in the first edition of his "Odyssey" (cf. Bien 2015: 213–214).

[4] Vlada Baranova (2017: 106) cites the request of the speakers of Azov Greek (documented in 1929) not to teach their children in Greek because they could not understand it.

[5] Before that they officially remained subjects of the Ottoman state.

2.2. Linguistic problems

In spite of multiple descriptions of Pontic, there still are many mysteries especially in case of Pontic outside Turkey, and Soviet and Post-Soviet Pontic is no exception here. In the next few subsections, I shall give a brief outline of the most problematic issues.

2.2.1. Subdialectal division

Even the first papers on Pontic acknowledged the existence of several local varieties in Asia Minor which differ on various linguistic levels (some details can be found in Papadopoulos 1958: ϛ–θ, ιϛ). It seems that in Soviet and Post-Soviet Pontic, one could also distinguish some local subdialects: for example, in Pontic of Batumi (Georgia) feminine nouns in plural nominative and accusative have flexion -es, while in Vityazevo (Southern Russia) and Iraga (Georgia) flexion -as is used instead (Eloeva 2004: 106). My own field research demonstrates some phonetic differences between Pontic of Abkhazia (1a) and Pontic of Southern Russia (1b):

(1) a. /foɣúme/ — informant female, born in 1966 in Gvarda/Lindava (Abkhazia), ancestors are from Ordu

  b. /fovúme/ — informant male, born in 1967 in Yessentuki (Southern Russia), ancestors are from Trabzon

Definitely, possible differences require special research which should consist of the following case studies:

  (i) regularity and levels of variation;
 (ii) number of local variants;
(iii) correspondence with local variants of Pontic in Asia Minor, i. e., it is important to find out whether Pontic subdialects of the USSR are "inherited" (brought from Pontos) or appeared after the dispersion of Pontic Greek in the "new" territory;
(iv) comparison with Pontic of Greece, Cyprus, Australia and other countries where Pontic is still spoken[6].

---

[6] There is no doubt that all variants of Pontic are mutually comprehensible but according to my interviewees from Southern Russia, there are certain differences in vocabulary, phonetics and morphology.

2.2.2. Language contacts

Quite a lot is already known about language contacts in Asia Minor but Soviet and Post-Soviet Pontic has hardly ever been analyzed from this point of view. Thanks to my former postgraduate student Liudmila Goroshilova, I have at my disposal several modern narratives from Georgia[7]. In these texts it is possible to trace some interference with Russian. The speakers do not use many lexical borrowings but prefer to incorporate Russian names of months although they know Greek names as well (*septémvrio* 'September', *oxtóvrio* 'October', *novémvrio* 'November' and *ðekémvrio* 'December')[8].

(2) *e-pín-ane*   *os*   *to*   ***maj   janvár^j   fevrál^j   mart   aprel^j***
     PST-do.IPFV-PST.3PL   till   DEF.N.SG.NGEN   May   January   February   March   April

'They were doing [it] until May, January, February, March, April'

Possibly, the phonetic phenomenon in (3) can be also treated as showing the impact of Russian because Russian phonetics lacks the sound /θ/:

(3) *tóra*   ***sa=léγ-o***
     now   FUT=speak.IPFV-PRS.1SG

'Now I shall tell [you about…]'

Surprisingly, in the speech of Pontic Greeks from Georgia, the interference with Georgian is extremely sporadic (Berikashvili 2022: 8) even in vocabulary, despite the fact that all are speakers are fluent in Georgian from a very young age. It is possible that we just lack the examples at our disposal, but, certainly, the matter requires further investigation.

2.2.3. Pontic and Azov Greek

In his famous book about the dialect of Lesbos Paul Kretschmer (1905: 18) suggested that Azov Greek should be one of the Pontic dialects because they both have palatalization:

(4)   Pontic *ʃér*[*in*] 'hand' and Azov Greek *ʃer^j* vs. Modern Greek χέρι

---

[7] These are narratives of elderly female speakers who are descendants of the immigrants from Trabzon.

[8] These could be, however, "new" borrowings from Modern Greek, since all these speakers regularly visit Greece and live with their "Greek-speaking" relatives.

Charalambos Symeonidis and Dimitrios Tombaidis support Kretschmer's hypothesis and point out that Azov Greek has many common features with Pontic like verb flexions, negations, particles, some lexemes and even idiomatic expressions (Symeonidis & Tompaidis 1999: 133–139). However, this hypothesis cannot be accepted without thorough discussion. It should not be forgotten that according to an important lexical isogloss, Pontic and Azov Greek belong to different groups — Pontic is an *índa*-type dialect, while Azov Greek is a *ti*-type (see also Kisilier 2022: 10–11).

3. "Pontic project"

The problems mentioned in section 2 cannot be resolved without a systematic multilateral approach which is impossible without representative collection of relevant and reliable data. Modern digital technologies allow to create a database with multiple options like classification, comparison, data representation and machine learning. The creation of such database is the main goal of so-called "Pontic project".

3.1. Digital collection

The project was launched several years ago when it became evident that the National Library of Russia possesses a collection of Greek books, journals and newspapers published the USSR between 1929 and 1937. The absence of any catalogue for this collection made it difficult to estimate its volume. Thanks to the efforts of Ilya Magin and participation of Yuri Polokhin it became possible to start the process of digitalization and to create online library[9].

The collection includes not only books, journals and newspapers in Pontic but also in Azov Greek and Demotic Greek published in the USSR. It was decided to digitalize everything regardless the dialectal attribution, and there are at least 3 reasons for that:

(i) sometimes it is not so evident which variant of Greek is used — dialectal or Demotic: the text may have both Demotic and dialectal words and grammatical forms;

---

[9] This library is available at: https://archive.org/details/sov-greek (accessed on 13.07.2022).

(ii) it is evident that if non-Pontic texts are excluded and remain undigitized, there will be no chance to start another Project involving these texts, so their existence may remain unknown;

(iii) there is a very tempting idea to identify the lexemes, grammar forms and possibly whole texts using algorithms based on machine learning.

In the future, we intend to widen our digital library by adding other texts in Pontic and Azov Greek including that were collected during the field research [10]. However, it is important to mention the problems which are most sure to emerge with the progress of the Project. Most digitalized texts still require recognition. The pages are faded, and not printed all that well; some are dark yellow or brown and the text is dark grey. So automatic recognition is difficult and likely to be compromised. Even a more difficult issue is a recognition of dialectal texts published in Greece where traditionally Greek letters with diacritics are used. These characters are absent from Unicode, and systems like ABBYY FineReader do not easily cope with them.

3.2. Dialectal database

Initially, there was an idea to create online dictionary of Pontic combined with a language corpus and sound recordings. We intended to create interface and translation in 4 languages (Russian, English, Modern Greek and Turkish), and the example of a dictionary entry in English may be found in (5) [11].

(5) **δουκάλιν** /ðuˈkalin/ Giresun, **δουκάλ'** /ðuˈkalʲ/ Argyroupoli, Dumanlı, Ordu, Trabzon (Papadopoulos 1958: 274), **δουκάλι** /ðuˈkali/ Trabzon (Tursun 2019: 236), **δουκάλ** /ðuˈkal/ noun, neuter
[ᴇᴛʏᴍᴏʟᴏɢʏ: Lat *ducalium* 'bridle']
ᴛᴏᴏʟs bridle, halter
Με το **δουκάλ'** λαλεί ατόν' '[s/he] manipulates him' (Papadopoulos 1958: 274; lit. [s/]he speaks with him by means of bridle)

---

[10] The digital library is free and open and we shall eagerly accept any text in Pontic or Azov Greek for publication.

[11] This work would have been impossible without assistance of the Pontic poet Viktor Stoforandov.

Δίγω τ' άλογον δουκάλιν '[I] loosen the reins and let the horse run' (Papadopoulos 1958: 274; lit. [I] give bridle to the horse)
DERIVATIVE *δουκαλεύω*

At the stage of the demo version, it became clear that there will be no funding and the ideology of the Project should be changed. Instead of a dictionary, it was decided to create a database using TLex software. This database primarily works with materials collected during field research but it can also include data from literary sources and linguistic descriptions (Figure 1).
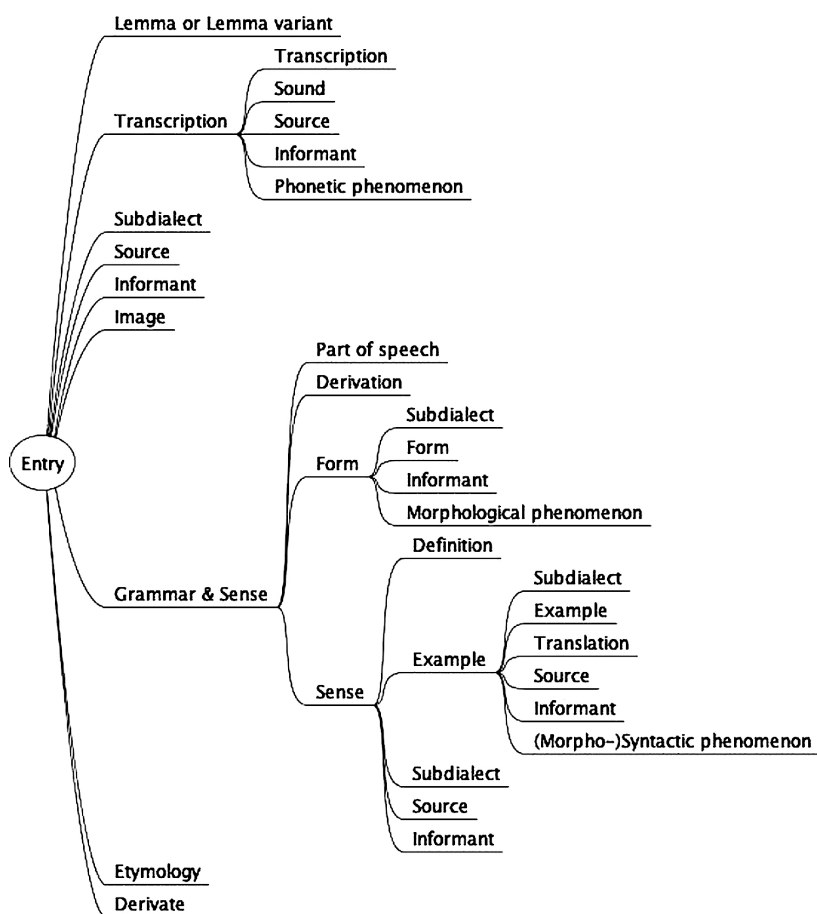

Figure 1: Database entry structure

All data within the entry is supported with multiple tags (metadata) describing the source and/or the interviewee/author of the text. This metadata will be helpful both for navigation and for various types of research like diachronic or gender studies. It is also possible to tag (partly manually, partly via an automated program) particular phonetic, morphological or even syntactic phenomena and then to sort them out according to a required pattern. The database exists in XML format.

On the next stage which in fact starts almost simultaneously with the first one, the database is connected with the digital collection of texts using readymade free software libraries adopted for the data involved. As a result, it will be possible to create a grammar analyzer using machine learning algorithms. All forms generated by means of the grammar analyzer will be automatically checked in the collection of digital texts. The form that is not found in any text will not be displayed within the paradigm until some new text is added where the form exists. We expect that adoption of array programming will automate the search of synonyms and antonyms and allocation of lexical and semantical classes.

The software libraries selected for adoption were mostly created for Standard Modern Greek and there is a chance that during the process of adoption new previously unknown peculiarities will become apparent. If so, using a software library created for Standard Modern Greek may become a new approach to comparison of Modern Greek dialects.

4. Conclusions and prospects

In comparison with a dialectal corpus, a dialectal database even in a very incomplete form may produce good results, make use of machine learning technologies and array algorithms and is able to function with a limited set of texts. Almost at any stage of its elaboration, the database may be converted into online dictionary or dialectal atlas and any update of the database will almost automatically result in the dictionary and atlas.

Creation of similar databases for various dialects offers amazing prospects for all kinds of comparative analysis and studies in dialectrometry. In order to test these possibilities, simultaneously with the Pontic database, I intend to start working on Azov Greek and Tsakonian databases.

References

Animitsa, G. A. & M. L. Kisilier. 2009. Istoriia Urumeev. Khronologiia osnovnykh sobytiĭ (1771–2003) [History of Urumejs. Chronology of main events (1771–2003)] Ch. 1. In M. L. Kisilier (ed.), *Iazyk i kul'tura mariupol'skikh grekov* [*Language and Culture of Mariupolis Greeks*]. Vol. I. *Lingvisticheskaia i etnokulturnaia situatsiia v grecheskikh selakh Priazov'ia. Po materialam ekspeditsiĭ 2001–2004 godov* [*Language and Ethno-Cultural Situation in Greek Villages of Azov Region. Field Research Data Collected in 2001–2004*]. Saint Petersburg: Aleteia, 25–64 [in Russian].

Arkhangelskiy, T. A. & M. L. Kisilier. 2018. Korpusa grecheskogo iazyka: dostizheniia, tseli i zadachi [Corpora of Modern Greek: achievements and goals]. *Indoevropeĭskaia i klassicheskaia filologiia* [*Indo-European Linguistics and Classical Philology*] 12(1)|: 50–59 [in Russian].

Baranova, V. 2017. Local language planners in the context of early Soviet language policy: the case of Mariupol Greeks. *Revue des études slaves* 88(1–2): 97–112.

Berikashvili, S. 2022. Contact-induced change in the domain of grammatical gender in Pontic Greek spoken in Georgia. *Languages* 7(2): 79. Available at: https://www.mdpi.com/2226-471X/7/2/79 (accessed on 10.07.2022).

Bien, P. 2015. *Kazantzakis and Linguistic Revolution in Greek Literature*. Princeton; New Jersey: Princeton University Press (Princeton Essays in European and Comparative Literature 6).

CGT — Corpus of Greek texts. Available at: http://www.sek.edu.gr (accessed on 13.12.2021).

CMG — Corpus of Modern Greek. Available at: http://web-corpora.net/GreekCorpus/search/?interface_language=en (accessed on 14.12.2021).

CSG — Corpus of Spoken Greek of the Institute of Modern Greek Studies. Available at: http://corpus-ins.lit.auth.gr/corpus/index.html (accessed on 13.12.2021).

Der Matossian, B. 2009. The Pontic Armenian communities in the nineteenth century. Ch. 10. In R. G. Hovannisian (ed.), *Armenian Pontus. The Trebizond-Black Sea Communities*. Los Angeles: Mazda Publishers, 217–244.

Eloeva, F. A. 2004. *Pontiĭskiĭ dialect v sinkhronii i diakhronii* [*The Pontic Dialect in Synchrony and Diachrony*]. Saint Petersburg: Saint Petersburg State University [in Russian].

Goutsos, D. 2010. The corpus of Greek texts: a reference corpus for Modern Greek. *Corpora* 5(1): 29–44.

HNC — Hellenic National Corpus. Available at: http://hnc.ilsp.gr (accessed on 10.12.2021).

Kachalov, Kh. 1934. *Καταλογος βιβλιον* [*Catalogue of Books*]. Rostov-on-Don: «Komounistis» [in Greek and Russian].

Kisilier, M. 2022. Contact phenomena in Azov Greek. *Languages* 7(2): 79. Available at: https://www.mdpi.com/2226-471X/7/3/174 (accessed on 8.07.2022).

Konstantinidou, Z. & A. Kyridis. 2014. The use of Pontic dialect today. *Mediterranean Journal of Social Sciences* 5(8): 627–633.

Kretschmer, P. 1905. *Der heutige lesbische Dialect verglichen mit den übrigen nordgriechischen Mundarten*. Wien: Alfred Holder.

Moseley, Ch. 2007. General introduction. In Ch. Moseley (ed.), *Encyclopedia of the World's Endangered Languages*. London & New York: Routledge, vii–xvi.

Özkan, H. 2013. The Pontic Greek spoken by Muslims in the villages of Beşköy in the province of present-day Trabzon. *Byzantine and Modern Greek Studies* 37(1): 1–21.

Papadopoulos, A. A. 1958. *Ἱστορικὸν λεξικὸν τῆς ποντικῆς διαλέκτου* [*Historical Dictionary of Pontic*] 1: Α–Λ. Athens: Myrtidi [in Greek].

Pavlidou, Th.-S. (ed.). 2016. *Καταγράφοντας την ελληνική γλώσσα* [*Making a Record of the Greek Language*]. Thessaloniki: Institute of Modern Greek Studies [in Greek].

Ralli, A. (ed.). 2015. *Πρόγραμμα ΘΑΛΗΣ: Πόντος, Καππαδοκία, Αϊβαλί: στα χνάρια της Μικρασιατικής Ελληνικής* [*Program ThALIS: Pontos, Cappadocia, Aivali: the Tracks of Asia Minor Greek*]. Patras: Laboratory of Modern Greek dialects [in Greek].

Savvov, D. N. 1931. K probleme iazyka i pis'mennosti grekov SSSR [About the language and the writing system of the Greeks of the USSR]. *Kul'tura i pis'mennost' Vostoka* [*Culture and Written Language of the East*] 10: 65–75 [in Russian].

Schreiber, L. & I. Sitaridou. 2018. Assessing the sociolinguistic vitality of Istanbulite Romeyka: an attitudinal study. *Journal of Multilingual and Multicultural Development* 39(1). Available at: https://www.tandfonline.com/doi/full/10.1080/01434632.2017.1301944 (accessed on 26.10.2020).

Sitaridou, I. & M. Kaltsa. 2014. Contrastivity in Pontic Greek. *Lingua* 146: 1–27.

Sokolov, I. 1932. Mariupol'skie greki [Greeks of Mariupol]. *Trudy Instituta Slavianovedeniia Akademii Nauk SSSR* [*Transactions of Institute of Slavic Studies of the Academy of Sciences of the USSR*] 1: 297–317 [in Russian].

Symeonidis, Ch. P. & D. Tompaidis. 1999. *Η σημερινή Ελληνική διάλεκτος της Ουκρανίας (περιοχής Μαριούπολης)* [*Modern Greek dialect of Ukraine (region of Mariupol)*]. Athens: The Committee for Pontian Studies (Supplement to "Archeion Pontou" ["Pontian Archives"] 20) [in Greek].

TLG — Thesaurus Linguae Graecae. Available at: http://stephanus.tlg.uci.edu (accessed on 08.12.2021).

Topalidis, S.. 2018. *The Relocation of Greeks from Pontos*. Manuscript.

Tursun, V. 2019. *Romeika-Türkçe sözlük. Trabzon Rumcası*. İstanbul: Heyamola yayınlatı.

Vergeti, M. 1991. Pontic Greeks from Asia Minor and the Soviet Union: problems of integration in modern Greece. *Journal of Refugee Studies* 4(4): 382–394.