Interactive Visualization of Dialectal Lexis Perspective of Research Using the Example of Georgian Electronic Dialect Atlas

Marine Beridze[1], Zakharia Pourtskhvanidze[2], Lia Bakuradze[1], David Nadaraia[1]
[1]Javakhishvili State University, [2]Goethe-University Frankfurt/M
E-mail: marineberidze@yahoo.com, pourtskhvanidze@em.uni-frankfurt.de, lia.bakuradze@tsu.ge, david.nadaraia@gmail.com

Abstract

This article presents a report of the results on the current situation in the development of two projects. These are (1) "The Large Georgian Dialect Lexicographic Database and the Georgian Electronic Dialect Atlas" and (2) "A Georgian Language Island in a Trans-Ethnic Area (GLITEA)"[1]. In the first project, the lexicographical component of the Georgian dialect corpus will be expanded and visualized cartographically. The second project examines the dialect of the Georgian – Fereydanian – spoken in Iran by the descendants of about 100 thousand Georgians, who were forcibly evacuated from east Georgia to Iran by Shah Abbas I in the period 1614 to 1616. The dialect is a typical case of a language island and offers the possibility for diverse linguistic research into language history, language contacts and language migration.

Keywords: dialectology, Linguistic Geography, dialectological lexicography, canonical visualization of linguistic data.

1 Introduction

The eighteen dialects of Georgian, three of which are spoken outside of Georgia, have been the subject of scientific research for about a hundred years. In the course of this period, the dialects were described by empirical and field research methods at the levels of grammar and vocabulary. In the 1930s, several documentations of the dialects were carried out by field research (Beridze V. 1938). As early as 1956, the Chrestomathy of the Georgian Dialects with Dictionaries (Dzidziguri (1956) was founded and published, then the Chrestomathy of the Georgian Dialects (Gigineishvili 1961). These created a whole series of monographs of the grammatical structure of the dialects (e.g. Jorbenadze 1988, 1989; 1991;1995;1998; Dzotsenidze 1974; Nizharadze 1975; Glonti 1975; Gachechiladze 1976; Meskhishvili 1981; Martirosov 1985; Gambashidze 1988; Chincharauli 2005; Tsotsanidze 2012). In recent years, the research groups have also published essays on the morphology of Georgian dialects (Gogolashvili 2017).

In this research tradition the dialectal phenomena were always considered within the framework of an adopted dialect standard and historically developed geographical boundaries. The migrations or the language contact phenomena were not taken into account at all. In the 1980s the basic idea of the Dialectal Atlas of the Georgian was conceived. In this context, a questionnaire was standardized and used in field research. In 2003 the project The Linguistic Portrait of Georgia was started. This project foresees the comprehensive documentation of the linguistic situation in Georgia and continues to this day. The Georgian Dialect Corpus is a result of this project, and forms the modern methodological

1 Supported by Shota Rustaveli National Science Foundation (SRNSF) [grant numbers 217008/217438].

approach to the study of the vernaculars. The structure of the corpus and the associated lexicographical database contains information on the geographical distribution of the data. Their visualization is the current phase of the project.

2 The Specifics of the Geography of Georgian Dialects – Historical and Linguistic Tradition

Georgia is composed of historically formed provinces, distinguished by peculiar cultural, ethnographic and linguistic characteristics (Figure 1, Figure 2.), with historical information about the dialectal diversity of Georgians available in the literature (Jorbenadze 1989; Sardjveladze 1975).

Figure 1. Ethno-culturally defined provinces of Georgia.

Figure 2. Geographical distribution areas of the dialects.

The Georgian dialectology, as a subject, was based on the historically and scientifically developed principle of the ethno-cultural classification of the country. This means that the geographical distribution areas of the dialects coincided precisely with the historically accepted limits of the distribution of ethno-culturally defined provinces of the country

3 Dialects in the Context of Migration

The migration processes in Georgia were constantly taking place with varying intensity. These were both ecologically and economically dependent and forced migrations through wars and raids. The traces of migration are still visible in the onomastics of the various localities. There were some great migration waves that have significantly changed the dialectological image.

- Massive expulsion in the 17th century from eastern Georgia to Iran near Esfahan by Shah Abbas and the emergence of the Fereydanian.
- Territorial redistribution in border areas to Azerbaijan in the 20th century and Turkey in the 19th century. Internal migrations for economic, environmental or legal reasons, occurred in the first half of the 20th century.
- Internal migrations resulted in compact settlements, but also dispersed branches.

If one considers the idea of the historical boundaries of the ethno-cultural provinces of Georgia in the context of migrations, then the picture shifts. The rigid borders are softened and small dialectal islands are created elsewhere. The dialect islands have their own language development, independent of the “mother dialect”.

4 The Status of the Current Dialectal Geography

The Georgian dialects, which through migrations created a new geographical image, developed in Turkey, Azerbaijan and Iran surrounded by completely different cultures, ethnicities and languages. This composition allows the research field to be considered as a large language laboratory by examining the dynamics of dialectal change, as well as language contact phenomena. The same applies to the dialectal islands within Georgia. For example, the Imeretian dialect, which is historically located in western Georgia, is also compactly represented in other areas of Georgia.

1. In Kakhetia (Lagodekhi, east Georgia) (Figure 4). Created by economic migration around 1905.
2. In Samtskhe-Javakheti (south Georgia) (Figure 3). Created by the tight resettlement under the Stalinist repression.
3. In Marneuli (south-east Georgia). Created by the so-called planned resettlement.

On the other hand, if you look at the Kakhetian dialect that is historically and solely located in eastern Georgia, you will find that there are at least five compact dialectal islands in this area: Imeretian, Ratchan, Pshavian, Khevsurian, Tushetian. A similar situation is also seen in Samtskhe-Javakheti. These dialectal "insertions" have different ages and degrees of isolation. Accordingly, they show different dynamics of language development.

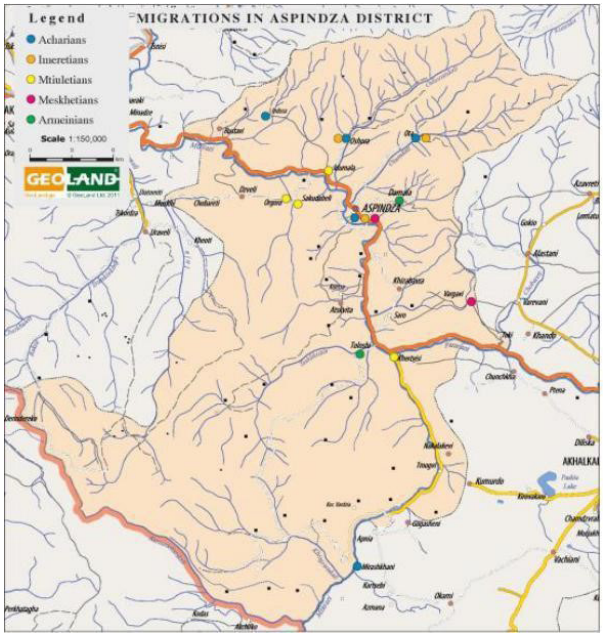

Figure 3. Dialectal islands in Aspinza.

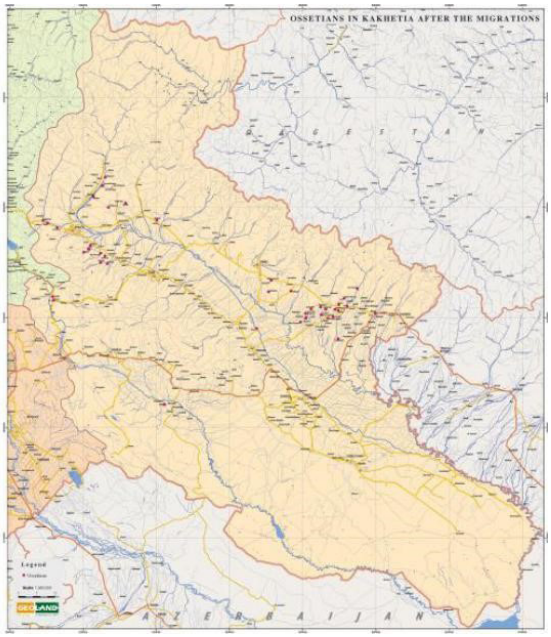

Figure 4. Language islands in Kahetia.

The dialectal distribution of the lexeme 'child' *bavšvi / bovši / balġi / boši* illustrates this observation (Figure 5).

According the fixed ethnographic boundaries the dialectal distribution of the lexeme 'child' is as follows:

balġi / bavšvi - Mtiuletian-Gudamaqrian, Pshavian, Tushetian, Khevsurian, Mokhevian (*bavšvi* is secondary lexical unit).

- *bavšvi / balġi* - Kakhetian, Kartlian (balġi is secondury lexical unit due to migrations from mountains).
- *bavšvi* - Javakhian.

- *bovši* - Imeretian.
- *boši* - Lechkhumian, Rachan.
- *baġvi* - Adjarian.
- *baġana / baġane* - Gurian.

On a strict search of the dialectal form bavšvi in the corpus texts, we will find exclusively Javakhian texts, but if we search the same lexeme based on place of elicitation, then we get at list four dialectal variants of bavšvi at the territory of Javakheti (Figure 6): *baġvi / balġi / bovši / boši.*

Figure 5. The dialectal distribution of the lexeme 'child'.

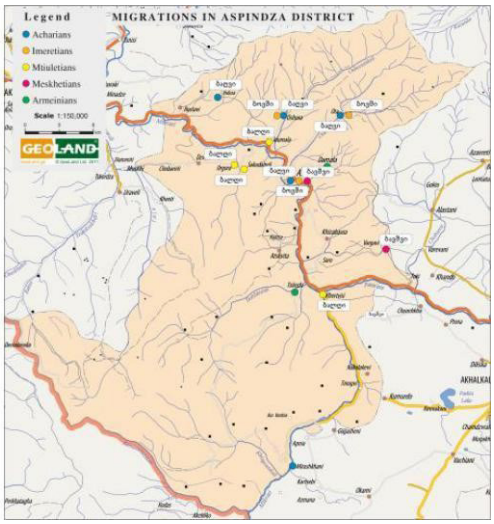

Figure 6. Dialectal variants of the lexeme 'child' at the territory of Aspindza (Samckhe-Javakheti)

5 Basic Question of the Canonical Visualization of Dialects

With regard to the conception visualization, we operate with the terms "operative imagery" and "recognizable vision" (both terms from Krämer 2009:94-123). For this we included the idea of cognition-winning visualization. Our goal is not simply to describe the migration processes, but to understand the reasons for migration in the geographical context. To achieve this goal, the documented linguistic data is linked to the geographical information of the migration. In the end, it should be visualized not only with the migration of the ethnic groups, but especially the migration of the dialects. The added value of this method lies in the justification of the geo-referentiality for the Georgian dialects. The expectation is to get a new perspective and thus new knowledge about the known language data and processes. On the one hand, it is the simple digital mapping of migration processes wherein the final product in the form of an interactive map has the explanative effect.

6 Traditional Solutions and New Tasks

The idea for the Georgian dialect atlas came for the first time from one of the founding fathers of the Georgian dialectology, Varlam Topuria, in the middle of the last century. In the 1980s questionnaires were drawn up and standardized in order to operate uniform field research and so create the language atlases of the individual linguistic regions (Kiziria 1984). The resulting questionnaire included about 1,000 questions on the basic levels of grammar. The work was interrupted in the 1990s and only

continued in 2006. However, the methodological battery of research was extended with digital instruments and the earlier, analog approach was replaced. Large language databases and complex search systems currently determine the documentation of Georgian dialects and their exploration. The use of digital geo-referentiality contours and the creation of dialectal maps offers additional ways of obtained knowledge from this research.

7 The Project: Linguistic Portrait of Georgia: Features and Details

As part of the project, the old, analog-documented data were systematized and digitized. Records received on magnetic tapes were transferred and secured in the new data carriers. In addition, new data was also collected and archived in a standardized process. The foundation of the dialect corpus was thus laid. At present, the Georgian Dialect Corpus is characterized by the follow statistical features:

- The lexicographic base was made using documented data in approximately 800 loci.
- The size of the corpus is 2,041,830 tokens.
- The number of the types is 460,631word forms.
- The number of the texts is 3,356.
- The number of the documented dialects is 18.
- The number of the annotated tokens is 425,631
- The word lists in the testing procedures contain 346,670 word forms.
- The number of all lexicographic articles is 102,638.
- The number of the published lexicographic articles is 54,000.

The dialect corpus and lexicographic base have a particular system of annotation resp. tagging. The common hierarchical structure for the marking of the grammatical and lexical features was constructed. The first stage contains the POS tagging and some parameters (like word fragments, affixes and so on), which are necessary in the tagging process. The second stage marks prepositions, particles of word formation and enclitics. In the last stage the marking system analyzes the semantic features like borrowing, terminology or idiomatic expressions. The annotation tag set is based on the Leipzig Glossing Rules and contains additional specific tags like ‘Fseudo’ for ‘pseudo standard language’, ‘Fragment’ for a ‘word fragment’ and ‘NonGeo’ for a ‘borrowed word’.

8 Dialectography of the Language Islands

The central point of the current phase of the project is the dialectography of the language islands. There is a special scientific interest in phenomena such as linguistic innovation, sub-dialectal forms, pseudo-literary lexemes, borrowings of all kinds, semantic shifts, idiomatic printouts obtained in the use of missing words, and so on. The Georgian Dialect Corpus incorporates the lexicon of the Georgian dialects and Laz in Turkey (corresponding to over nine and six thousand words and articles, respectively), in Iran – Fereydanian (with over six thousand words and articles) and in Azerbaijan – Ingilo (with over 11 thousand words and articles). In field research on the language islands, a little-known phenomenon of the transformation of the language assistants (informants) was observed with regard to independent scientific research on their own dialects. Over time, a special perception of one’s own language isolation is created, and an analytical-structural approach is established. It is thus necessary to later get a naïve linguist to help with the elicitation of the island language. The corpus contains the editing tool “lexicographical editing”, which allows the processing of a lexical registration at different levels (Beridze M. 2017). This is especially important in the description of

language islands, because there are many words initiated by the contact to the standard language of the motherland. A particular difficulty is the description of the internal migrations on the language island, when no certificates exist. An attempt to study the internal differentiation of the Fereydanian dialect was done with the application of dialectometry. The empirical data elicited from the seven Fereydanian villages resulted in a Levenshtein distance matrix (Table 1), which was visualized (plotted) differently.

Table 1. Levenshtein Distance Matrix

	DASH-KASAN	AGCHE	BOIN	NEHZA-D_A	SIBAK	FEREY-DUN_S	MIAN-DASHT
DASHKASAN	0.0	2.142	2.054	2.666	2.405	2.192	2.644
AGCHE	2.142	0.0	2.888	2.567	1.916	1.607	2.533
BOIN	2.054	2.888	0.0	2.964	2.733	3.0	2.666
NEHZAD_A	2.666	2.567	2.964	0.0	1.677	2.3	2.914
SIBAK	2.405	1.916	2.73	1.677	0.0	2.325	2.882
FEREYDUN_S	2.192	1.607	3.0	2.3	2.325	0.0	2.469
MIANDASHT	2.64	2.533	2.666	2.914	2.882	2.469	0.0

The tree diagrams of the language distances were based on the corresponding geographical data (mapped polygons) and the tendency of the internal grouping was designed (Figure 7).

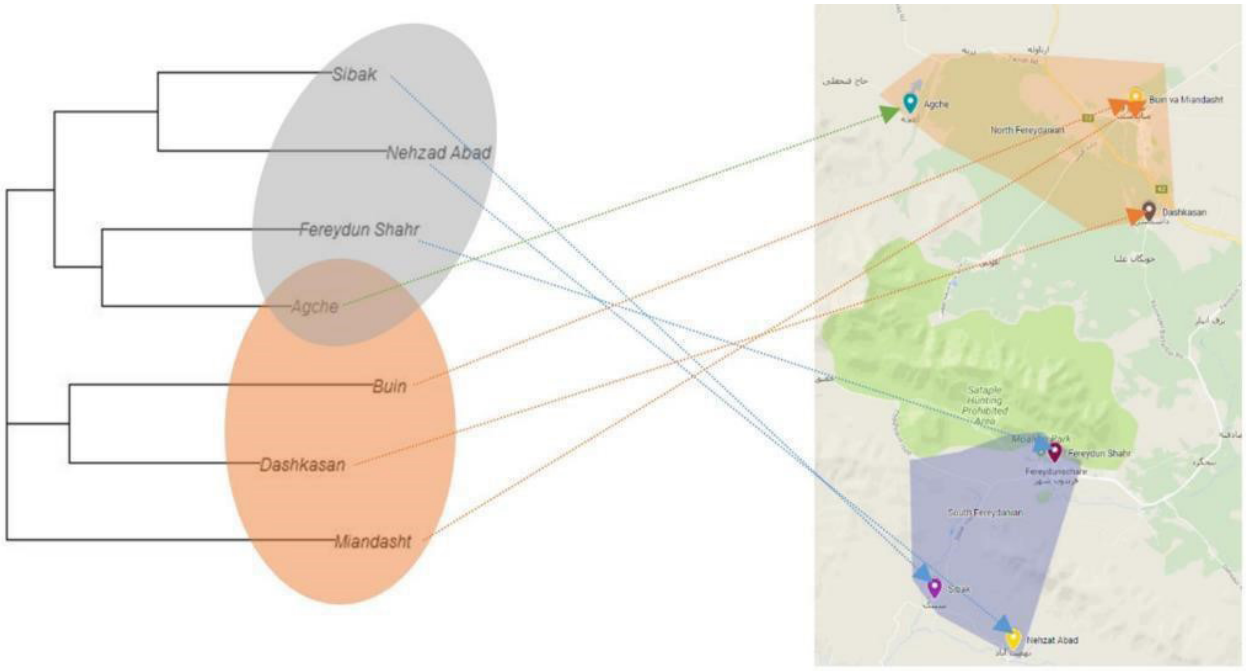

Figure 7. Geographically compact parts of the language islands also show linguistic proximity to each other, whereby a settlement (Agche) occupies the middle position, which has to do with its history of origin by immigrants from the south-Fereydan villages.

In the traditional dialectology in relation to the Fereydan, it was always assumed that the Georgians forced into Iran in the 17th century were the speakers of a certain dialect, called Kakhetian, and therefore the current linguistic situation on the language island should prove a common variant. This assumption was based mainly on the historical scientific evidence. The results of the dialectometry

and the different visualizations have clearly shown that the Fereydanian language island is not a monolithic language unit, but is experiencing internal migrations and shifts in language development. However, the empirically justified dialectal diversity of this language island is based on the quantifiable linguistic method of the dialectometry

9 Aggregate for Geo-referential Visualization

A central point of the projects is the establishment of the digital dialectal atlas of the Georgian. The atlas is based on the data from the Georgian Dialect Corpus and is designed using an aggregate of geo-referential visualization, and with the use of this in both projects a complex link between the corpus and the possibilities of digital cartography is understood. The text database, the lexicographical database and the geo-referencing algorithm interlock in the way that the geographical distribution of dialectal phenomena is more precisely and spatially represented empirically.

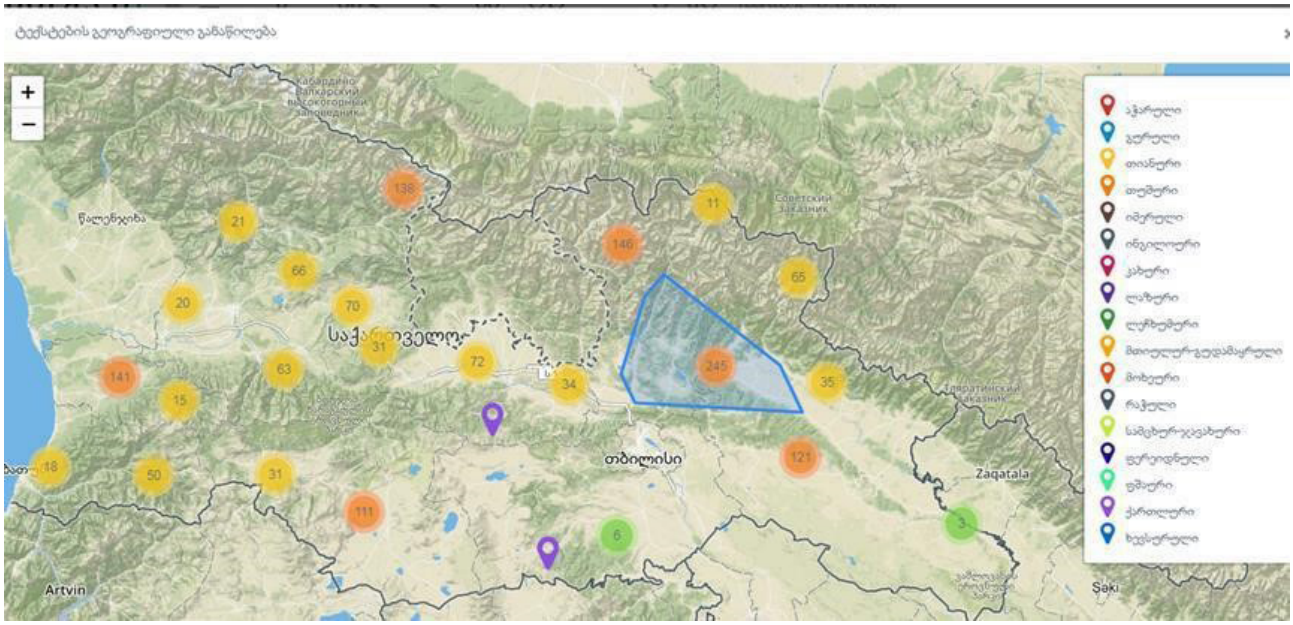

Figure 8. The points with numbers show the number of documented texts and the geographical areas can be mapped as polygons

The focus of the aggregate is the database of digitized dialectal texts, which are provided with detailed metadata. All texts archived in the Georgian Dialect Corpus contain a reference to the geographical location where they were documented. When the text database is tokenized, the meta-information of the entire text is projected onto the individual tokens. The metadata acts as the parent information that is inherited by the child tokens. At the end of this process, each token gets an additional geo-referential write-up that can be visualized. With the automatic lexicon entry of a token, the geo-referential attribution is included and is part of the lexicon article. Thus, the text database guarantees the exact location of tokens from the texts and lexical entries in the lexicographical database. The freely usable geodata for visualization are taken from OpenStreetMap and Geojson. By linking the geo-referential metadata of individual tokens with the information from the OpenStreetMap, a digital atlas is created that visualizes the distribution of dialectal phenomena accordingly. The collections were developed in the lexicographical database, and were created independently of the texts. These collections come from the research tradition of the middle and end of the last century, and are not always uniform in terms of meta-information. In most cases, however, the assessment of the location is the correct one. The entire database is annotated on several levels that are hierarchically structured. At the first level, POS tagging takes place. At the second level, the specification of the grammatical information

includes categorical properties of the analyzed elements. The further level of annotation sets the semantic properties of the tokens. The lexical features are supplemented with additional information and corresponding tags.

10 Prospect

The main challenge for the future remains the construction of a sub-corpus, which contains only the language data of the migrated dialects. This means the geo-referential documentation of the Georgian dialects in Turkey, Iran and Azerbaijan, and the corresponding material is currently being prepared. The visualization of the data in the context of the geo-referentiality presents us with the task of constructing a corresponding multi-view-mask, which will be web-based.

References

Beridze, M. et al. (2015) Dialect Dictionaries in the Georgian Dialect Corpus, Logic, Language, and Computation/ XIV Springer. Pp. 82–96.

Beridze, M. et. al. (2017) Georgian Dialect Corpus: Linguistic and Encyclopedic Information in Online Dictionaries. In *Journal of Linguistics/Jazykovedný casopis. The Journal of Ludovít Štúr Institute of Linguistics, SAV. N.68/2*. Pp. 109-121.

Beridze, M. et al. (2009) The Corpus of Georgian Dialects, In *Proceedings of the NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference Smolenice, Slovakia. Tribun.* Pp. 25-35.

Beridze, M. et al. (2015) The Georgian Dialect Corpus: Problems and prospects. In *Proceedings of the conference on Historical Corpora Challenges and Perspectives, Frankfurt.* Pp. 323–333.

Beridze, M. et al. (2011) Dictionary as a textual component of Corpus (Georgian Dialect Corpus). In *Proceedings of the conference on corpus linguistics, St. Petersburg.* Pp. 92-97.

Beridze, M. et al. (2014) Lexicographical concept of Georgian Dialect Corpus and problems of morphological analysis.).

In *Proceedings of the conference on Applied Linguistics in Science and Education, Knijnii dom, St. Peterburg*. Pp. 91-94.

Beridze, M. et al. (2016) Lexicographic Potential of the Georgian Dialect Corpus. In *Proceedings of the XVII EURALEX International Congress, Lexicography and Linguistic Diversity.* Pp. 300-309.

Beridze, V. (1938) Vocabulary of the Kartvelian Languades, I, 164 p.

Chincharauli, Al. (2005) *Khevsurian Dictionary*, 1177 p.

Dzidziguri, S. (1956) *Chrestomty of the Georgian Dialects with Dictionaries*. Tbilisi, 401 p.

Dzotsenidze, K. (1974) *Upper Imeretian Dictionary*, 645 p.

Explanatory dictionary of the Georgian language (1950-1964) 8 vols. Tbilisi: Georgian Academy of of Sciences.

Gachechiladze, P. (1976) *Lexical material of the Imeretian dialect*, 182 p.

Gambashidze, R. (1988) *Dictionary of Ingiloan dialect of Georgian Language*, 629 p.

Gigineishvili, I. et al. (1961) *Georgian dialectology*, I, 732 p.

Glonti, Al. (1975) *Dictionary of Georgian dialects*, 411 p.

Gogolashvili. G. et al. (2016) *Morphology of the Contemporary Georgian Language: Dialects*, II, 916 p.

Jorbenadze, B. (1989) *Georgian Dialectology*, I, 636 p.

Jorbenadze, B. (1998) *Georgian Dialectology*, II, 675 p.

Jorbenadze, B. (1991) *The Kartvelian Languages and Dialects*, 272 p.

Jorbenadze, B. (1995) *Dialects of the Kartvelian Languages*, 448 p.

Kiziria, A. et.al. (1984) *Questionnaire for Dialect Atlas Material*, 80 p.

Krämer, S. (2009) Operative Bildlichkeit. Von der Grammatologie zu einer “Diagrammatologi? Reflexion über erkennendes Sehen. In *Martina Heßler and Dieter Mersch (Eds.), Logik des Bildlichen. Zu Kritik der ikonischen Vernunft, Bielefeld: transcript, 2009.* Pp. 94-123.

Martirosov, A. (1985) The Main Issues of the Study of Georgian Dialect Vocabulary and Compilation of Dictionaries, In *Ibero-Caucasian linguistics, XXIII.* Pp. 139-148.

Meskhishvili, M. et al. (1981) *Dictionary of Kartlian Diallect*, 551 p.

Nizharadze, S. (1975) *Adjarian dialect*, 231 p.

Sarjvelasze, Z. (1975) *The issues of Georgian literary language history*, 271 p.

Speelman, D. and D. Geeraerts. (2008) 'The role of concept characteristics in lexical dialectometry', In *International Journal of Humanities and Arts Computing 2 (1–2).* Pp. 221–42.

Szmrecsanyi, B. (2008) 'Corpus-based dialectometry: aggregate morphosyntactic variability in British English dialects', In *International Journal of Humanities and Arts Computing 2 (1–2).* Pp. 279–96.

Szmrecsanyi, B. (2010) *The Morphosyntax of BrE Dialects in a Corpus-based Dialectometrical Perspective: Feature*

Extraction, Coding Protocols, Projections to Geography, Summary Statistics. Freiburg: University of Freiburg.

URN: urn:nbn:de:bsz:25-opus-73209. Available online at: http://www.freidok.uni-freiburg.de/Volltexte/7320/

Tsotsanidze, G. (2012) *Dictionary of Tushian Dialect*, 319 p.

Trudgill, P. (1974) 'Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography', In *Language in Society 3 (2).* Pp. 215–46.

Viereck, W., H. Ramisch, H. Händler, P. Hoffmann and W. Putschke (1991) *The Computer Developed Linguistic Atlas of England*. Tübingen: Niemeyer.

Georgian National Corpus http://gnc.gov.ge

Georgian Dialect Corpus http://corpora.co