Building a Dialectological Lexicological Database of Georgian Cognates for Digital Analysis[1]

Hélène Gérardin (Paris)

DOI: https://doi.org/10.62235/dk.3.2024.8511
helene.gerardin@inalco.fr || ORCID: 0009-0003-7406-3901

Abstract: This paper examines the challenges involved in creating a lexical database for Georgian dialects. It begins by outlining the methodological approaches to data collection and formatting, followed by an overview of the current version of the database, and its applications for linguistic analysis. Designed to facilitate a wide range of research, the database supports investigations such as Levenshtein distance calculations and diachronic and diatopic comparisons. The long-term goal of this project is to develop an open, accessible resource that can be gradually enriched with new data, advancing computational tools and deepening our understanding of Georgian and Kartvelian dialects.

Keywords: Georgian, Dialectology, Levenshtein distances, Diatopic and diachronic comparison

Introduction

The aim of this paper is to present a Digital Lexical Database of Georgian dialects that I have been building for a few years. The idea was initiated during the IDEX project Linguistic Dynamics of the Caucasus (LaDyCa),[2] which was conducted in 2017–2018 in collaboration of Sorbonne University (Paris) and Ilia State University (Tbilisi).[3] Throughout this project, I was in charge of collecting Georgian data (recording texts and lexical materials) with the intention of managing them via computer programs (e.g. Markov Models, Levenshtein distances). It was within this framework that the idea of focusing on lexical data emerged, leading to the decision to continue the processing to create a comprehensive database for broader purposes. The present article aims to describe the stages of this new task, along with its methodology, challenges and caveats.

After offering a brief overview of Georgian dialectology, I will delve into the issues, methods and challenges of my work. Following that, I will show the results of the created database, starting with the application of Levenshtein Distances to the database – a collaborative endeavour with colleagues from the LaDyCa project. In addition, I will present the current state of the materials and provide some examples of potential uses for the database. In the conclusion, I will attempt to delineate further steps necessary for advancing this research.

[1] I am grateful to Jost Gippert and Manana Tandaschwili for their interest in this work and the opportunity they provided me to present it at the conference "Digital Caucasiology – A Change of Paradigm" and subsequently publish it in this journal. I would also like to extend special thanks to Donald Rayfield and George B. Hewitt for their valuable feedback and encouragement.

[2] The IDEX projects ("Excellence Initiatives", in French "Initiatives d'excellence") are part of the "Investments for the Future" programs, which are initiatives set up by the State of France and aimed at creating multidisciplinary higher education and research institutions in the country that would be globally competitive. For details as to LaDyCa see Léonard 2019b.

[3] I thank Tamar Makharoblidze and Jean-Léo Léonard for giving me the opportunity to lay the first steps of the work presented here.

1. Overview of Georgian dialectology

Georgian exhibits a wide linguistic and geographical variation, resulting in the development of approximately 15–20 dialectal or subdialectal varieties. Several linguists have proposed different classifications,[4] but none of them has achieved unanimous acceptance. According to Jorbenadze 1989, Georgian dialects can be divided into two branches, along an East-West axis, which also corresponds to a historical and geographical boundary:

- Eastern Dialects
    - Mountain dialects: Pshavian, Khevsurian, Tushetian, Mokhevian, Mtiuletian-Gudamaqrian
    - Dialects of the plain: Kartlian, Kakhetian, Ingiloan, Fereydani
    - Meskhian, Javakhian
- Western Dialects
    - Ratchian
    - Imeretian, Letchkhumian
    - Gurian, Adjarian (+ Imerkhevian)

Standard Georgian is based on the Kartlian dialect, which is spoken in the area encompassing both the historical capital (Mtskheta) and the current capital (Tbilisi). Mutual intelligibility between Standard Georgian and the dialects is almost complete. All dialects are oral varieties with no writing tradition. They are seriously endangered, yet they offer data of primary importance for understanding the history and development of Georgian. As such, they constitute an important repository of grammatical and lexical categories that are not or no longer attested in Georgian. For this reason, their description, starting with a review of the lexicon, is urgent and promising.

2. Why a lexical dialectological database?

2.1. Aims

The current research aims to create a lexical database of cognates including items for all dialectal varieties, and later archiving the database online, facilitating its future expansion. Such work can be useful not only for archiving and processing dialectological materials but also for studying the linguistic variation of Georgian, as well as refining the classification of Georgian dialects. Furthermore, it is an efficient way of exploring and reconstructing the diachrony of the language and enhancing our understanding of the Kartvelian family.

This unique database is thus intended to be used for various subsequent studies and research frameworks.

[4] See for instance Chikobava 1952, Shanidze 1957, Dzidziguri 1970 and Jorbenadze 1989; for a cartographic representation and discussions about migrations, see Beridze et al. 2018.

2.2. Challenges

The heart of the work involves bringing together lexical units common for all Georgian dialects. This has two methodological implications: first, the need to eliminate borrowings (which are numerous, especially in dialects that are in contact with other languages); and secondly, the identification of words common enough to be encountered in corpora or elicited. Last but not least, it is important to compile a list extensive enough to be considered representative (with no less than a hundred items).

Due to these difficulties and the scarcity of dialectological resources, we must combine a diverse array of sources and compile both written and oral corpora.

2.3. The corpus

2.3.1 Written sources

Written sources primarily include dictionaries and word lists, usually found appended to text collections and descriptive works. A notable problem of such lists is their tendency to contain lexemes which deviate the most from the standard while common lexemes are often lacking.

Furthermore, collecting the items can also be carried out on the basis of published texts and online corpora.[5] A caveat when using texts is that lexemes used in context usually appear in modified grammatical forms. This requires determining the citation form.

Another common problem across all types of written sources is that transcription systems are not homogeneous (for instance, the sound [w] is transcribed as either *w*/ჳ or *v*/ვ or *u*/უ). This requires adapting the phonetic notation of certain items.

2.3.2. Oral sources

Oral corpus collection involves recording isolated words from speakers in the villages where the dialects are spoken. However, this task is far from being simple. First of all, the number of dialect speakers is dwindling and some areas remain difficult to access. Secondly, in the territory of Georgia, almost all dialect speakers are able (and used) to speak Standard Georgian, due to its spread through secondary education, the media, new communication tools, and recent migrations. As a result, speakers may not always be aware of the boundary between dialect and standard. However, dialects are still often associated in people's minds with inferior and non-prestigious ways of speaking. A consequence of this is that even individuals who only speak in a dialect automatically switch to standard Georgian as soon as they are recorded. Combining dialect and recording material is sometimes a real challenge for the linguist!

For all these reasons, in order to minimise interference with Standard Georgian, I had to develop an alternative elicitation protocol. I created a slideshow of pictures representing the target words, so that the speakers only had to mention what they saw in the pictures. Each slideshow comprises approximately 40 items, so that each inquiry lasts approximately 20–40 minutes.

[5] Electronic corpora are easily available, see for instance the Georgian Dialect Corpus (http://corpora.co) and the Georgian National Corpus (http://gnc.gov.ge), not forgetting the large database of TITUS (https://titus.uni-frankfurt.de/texte/texte2.htm#georgant).

Keeping word lists concise is essential, since speakers tend to get tired after 40 words within the session. The pictures should be easily identifiable, and the speakers are invited to label the items and avoid discussing them as much as possible in order to reduce interference from Standard Georgian. The screenshot in Fig. 1 presents an extract from the slideshow:

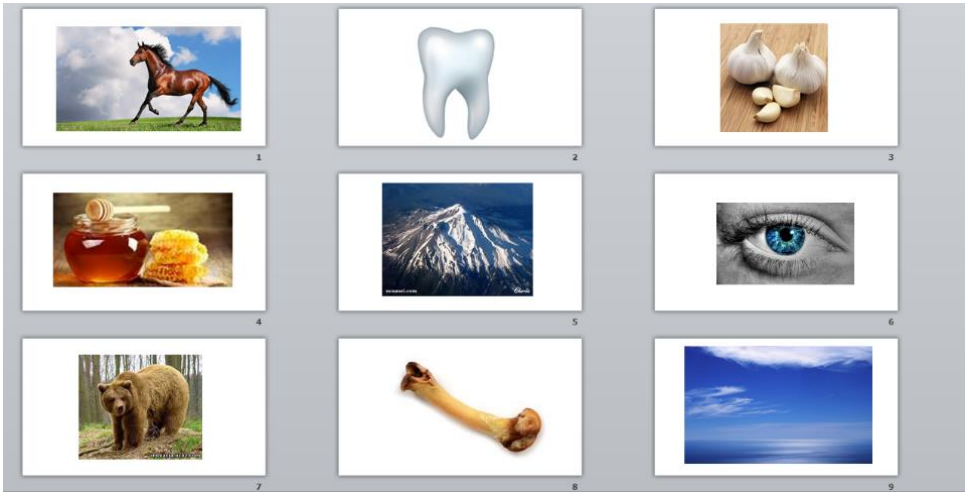

Fig. 1: Sample from the slideshow presenting lexical items

Such lexical elicitations can still pose a number of challenges. For instance, some lexemes cannot be conveyed through images, which requires seeking alternative methods.

I have tried to devise other ways for the speakers to produce the words as spontaneously as possible. While visual representations work well for numbers and colors, they prove difficult for qualifying adjectives, verbs or some nouns conveying functions or relations when not associated with any context. In this case, I had to communicate with the speakers in Standard Georgian to explain what the target term was. This gave rise to the use of various strategies, such as prompting people to guess words through completing sentences, engaging in logical enumerations, or even translating items from Russian or English. In any case, the experience gained from fieldwork in dialectology underlines the importance of avoiding pronouncing directly the Standard Georgian word because on hearing such forms, most speakers will automatically repeat them, making it impossible to capture the 'true' dialectal form. The screenshot in Fig. 2 presents a selection of such adapted slides taken from the end of the slideshow.

Fig. 2: Sample of adapted items

3. First results

3.1 The current version of the Database

Once collected, the materials must be transcribed to fill the database. The result is a table listing all the items classified by dialect and accompanied by their translation into English and French. The items are represented by their Standard Georgian lemma and organised in alphabetic order. To facilitate comparison and diachronic analysis, we have also added the forms in Old Georgian as well as reconstructed Kartvelian etymons when they are known (at the left). In addition, in order to broaden the base to include other Kartvelian languages, columns were created for Megrelian, Laz and Svan (positioned to the right of the Georgian dialects). The lexical Database, after formatting, exhibits the structure shown in Fig. 3 (extract).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	Etymo	Grg	Old Gr	Khev	Psh.	Tian.	Mokl	Mtiul.	Gud	Tush	Kartl.	Kakh.	Kiziq.	Fer	Ing.	Djav.	Samtsk	Ratch.	Im	Lech.	Adj.	Tao	Imerkh	Gur.	Mingr	Laz	Svan	Balzer	Lachk	Comi	Meanin
2	*kac	kaci	kaci	kaci	kaci	kaci	kaci	kaci	kaci	kacī		kaci	kaci	kaci	kac	kaci	kaci	kaci	kac	kaci	kaci	kaci	kaci	kaci	koči	koči	čäši, čaš			Svan	man
3	*kaḳal	kaḳali				kaḳali		kaḳali				kaḳali	kaḳali	kaḳa	kaḳal			kaḳali		kaḳali	kaḳa	kaḳali			kaḳal	kaḳal	kaḳ, gaḳ(a), kaḳra			Mingr	nut
4		kali	kali	kali	kali	kali	kali	kali	kali	kalī, k	kali	kali	kali		kal	kali	kali	kali, kal	kal	kali	kali	kali	kali	kali						Khev	women
5		karga	kargad	karga	karga	karga(d)	karga	karga(t)			karka	karga	karkat		[nama]	karga	kargat	karkat	kaat,	karka;	ka;	kāga	kärga								good
6		kari		kari	kari		kari	kari		karī		kari	kari		kari [kūl	kari	kari	kari	kari											Ingil.	wind
7		kargi	kargi	kargi	kargi, kai		kargi	karki, kai		karga	karki	kai, kı	kai, kı		[nan	kaj	kai, k	karki, ka	kaj,	karki	kaj		kaj	karka							good
8	*katan	katan	katami		katami		katan	katami				katam	katam		katan	katam	katami	katami	kat	katami			katami		kotom	kotur	katal				chicken
9		keipi		kaibi	keipi		keipi	keipi				keipi	keipi		keifī			keipi	keipi				kejpi								feast
10		kidev		kide(	kide	kider	kiden	kidav,	kida	kiden	kidev	kiden(	kide, l		kide	kede	kide	kide, ki	kid	kido,	kider	kide	kide,	kido							again
11	*klde	klde	klde	kilde	klde		klde	klde		kdej,	kilde	klde	klde		[daamana	klde	klde	kde, k(l)	kdı	kte,	kilde	kilde	klde	ktei	kyrde,	kirde	kylde, k(il)de				rock
12	*kmar	kmar	kmari	kmar	kma	kmar	kmar	kmari			kmar	kmari	kmari		kmari	kmari	kmari	kmari	kmari						komoi	komoǯ,	kimoǯ				husban
13		kvali		kvali	kvali		koli	koli	koli	kvalī		kvali	kvali		[rade pa]		kvali	kvali					kwali								trace
14		kveqana		koqa	xveqana,		xoqa	koqan	koqa	qveqanaj		kweqı	kweqı		[keš	[dūn	köqan	köqana	kveqana	koqana		koqana		kweqana						< *kv	world,
15		korćili		qorći	korćili		qorći	korćili	korćili			korćili	korćili		korç	korç	korćili	korćili	xorćili (l	korćili		dako	[duguni]								marriag
16	*ǩowz	kovzi		kozi			kovzi	kovzi		kozi	komz	kovzi	kovzi		kovzi		konzi	konzi, k	kozi (zm,	gl	kobzi			kob:	kowz,	kuz,	kiz				spoon
17	*kwa	kva	kwa	kva	kva		kva	kva		kvaj		kwa	kwa		kua	kvaj	kua	kua	kvı	kva	kva	kwa	kwa	kvai	kua, p	kwa,	pl : kwal				rock
18	*kverc	kvercxi		kvercxi			kverc	kvercxi			körcx	kverc	kverc		kerc	körc	körcxi	körcxi,	kvercxi			kvercxi	kwercxi							Kartv	egg
19	*ḳbil	ḳbili		ḳbili	ḳbili			kvili	kvili	kmilī		ḳbili	ḳbili		ḳbili			ḳbili		ḳbili,	ḳybil	kybil	ḳbili		ḳəbir,	ḳibi(r),	ḳibr, ḳirb,	ćibr			tooth
20		kvira		kvira	kvira		kvire	kvira,	kvir	kvira'ej		kvira	kvira		kvir	kūraj,	kūria	kvira	kvire				kwire								Khev	week
21		lamazi		lama	lama	lama	lama	lamazi	lama	lamazī		lamazi			lama	lamazi	lamazi	lamazi	lamazi				lamazi								Mokh	beautifi
22		loqa		[kuri]			laqva	laqva,	laqva,	laqo		loqa	loqa		loqa																cheek
23		magr	magran	magr	magram		magr	magra(m)		magra	magram				magr	mara	mara	mara, m	mara(m),		mara		mara	mara							but
24	*mama	mam	mamaj	mam	mama		mam	mama	mama			mama	mama	man	mamaj			mama	ma	mama	[baba,	babo]		man	muma	[mum	mū			Mingr	father
25		0	13	2	4	17	2	0	12	9	15	1	2	3	8	8	4	3	6	13	11	15	6	15	14	15	18	23	23	14	0

Fig. 3: Current database after formatting

The current version of the Database includes 243 items, representing approximately 5,000–6,000 items. Some fields still require completion.

3.2 Application of Levenshtein Distances

The first instance of putting this database to practical use involved the application of Levenshtein Distances. The procedure was conducted during the LaDyCa Project by Laure Picard, Jean-Léo Léonard and myself, using the Gephi software.[6] The results of this study were presented during the project[7] and subsequently partially published.[8]

According to Beijering, Gooskens & Heeringa (2008), the Levenshtein algorithm is a string-based distance measure that quantifies the differences between the (phonetical) shapes of corresponding words in different dialects or closely related languages. It calculates the minimal costs required to change a string of segments into another by means of insertions, deletions or substitutions. The resulting analysis of the Georgian data by Laure Picard is shown in the diagram in Fig. 4.

[6] The *Open Graph Viz Platform*, see https://gephi.org.
[7] Picard, Gérardin & Léonard 2018.
[8] Léonard 2019 and Léonard & Makharoblidze 2022.

	Grg	Khev.	Pch.	Mokh.	Mtiul.	Touche	Kartl.	Kakh.	Kiziq.	Ferey.	Ingil.	Meskh.	Ratch.	H-Imérét.	B-Imérét.	Letchkh.	Adj.	Tao	Imerkh.	H.-Gour.	B.-Gour.
Grg		0,123	0,093	0,15	0,122	0,305	0,127	0,079	0,093	0,183	0,518	0,104	0,119	0,162	0,19	0,25	0,225	0,239	0,176	0,264	0,255
Khev.	0,123		0,138	0,187	0,164	0,26	0,198	0,168	0,142	0,192	0,49	0,183	0,201	0,224	0,247	0,262	0,274	0,234	0,208	0,324	0,309
Pch.	0,093	0,138		0,147	0,116	0,283	0,144	0,105	0,088	0,178	0,478	0,137	0,151	0,177	0,212	0,261	0,241	0,257	0,165	0,312	0,304
Mokh.	0,15	0,187	0,147		0,109	0,35	0,196	0,174	0,164	0,172	0,442	0,198	0,223	0,253	0,25	0,313	0,318	0,282	0,216	0,363	0,343
Mtiul.	0,122	0,164	0,116	0,109		0,337	0,176	0,128	0,118	0,183	0,458	0,169	0,188	0,211	0,232	0,298	0,273	0,266	0,181	0,301	0,305
Touche	0,305	0,26	0,283	0,35	0,337		0,336	0,336	0,305	0,337	0,478	0,334	0,34	0,337	0,353	0,384	0,372	0,399	0,32	0,404	0,391
Kartl.	0,127	0,198	0,144	0,196	0,176	0,336		0,097	0,118	0,172	0,461	0,134	0,143	0,172	0,196	0,235	0,242	0,236	0,164	0,245	0,248
Kakh.	0,079	0,168	0,105	0,174	0,128	0,336	0,097		0,036	0,19	0,522	0,114	0,143	0,191	0,205	0,252	0,208	0,224	0,163	0,277	0,271
Kiziq.	0,093	0,142	0,088	0,164	0,118	0,305	0,118	0,036		0,172	0,504	0,14	0,16	0,205	0,213	0,259	0,226	0,207	0,143	0,291	0,283
Ferey.	0,183	0,192	0,178	0,172	0,183	0,337	0,172	0,19	0,172		0,436	0,181	0,202	0,237	0,254	0,339	0,265	0,265	0,224	0,353	0,337
Ingil.	0,518	0,49	0,478	0,442	0,458	0,478	0,461	0,522	0,504	0,436		0,495	0,511	0,507	0,504	0,566	0,513	0,519	0,528	0,545	0,537
Meskh.	0,104	0,183	0,137	0,198	0,169	0,334	0,134	0,114	0,14	0,181	0,495		0,137	0,173	0,202	0,276	0,261	0,289	0,215	0,29	0,288
Ratch.	0,119	0,201	0,151	0,223	0,188	0,34	0,143	0,143	0,16	0,202	0,511	0,137		0,144	0,168	0,19	0,221	0,273	0,204	0,234	0,241
H-Imérét.	0,162	0,224	0,177	0,253	0,211	0,337	0,172	0,191	0,205	0,237	0,507	0,173	0,144		0,125	0,219	0,257	0,275	0,222	0,229	0,235
B-Imérét.	0,19	0,247	0,212	0,25	0,232	0,353	0,196	0,205	0,213	0,254	0,504	0,202	0,168	0,125		0,175	0,249	0,288	0,217	0,2	0,192
Letchkh.	0,25	0,262	0,261	0,313	0,298	0,384	0,235	0,252	0,259	0,339	0,566	0,276	0,19	0,219	0,175		0,243	0,357	0,255	0,245	0,26
Adj.	0,225	0,274	0,241	0,318	0,273	0,372	0,242	0,208	0,226	0,265	0,513	0,261	0,221	0,257	0,249	0,243		0,23	0,226	0,288	0,294
Tao	0,239	0,234	0,257	0,282	0,266	0,399	0,236	0,224	0,207	0,265	0,519	0,289	0,273	0,275	0,288	0,357	0,23		0,182	0,301	0,323
Imerkh.	0,176	0,208	0,165	0,216	0,181	0,32	0,164	0,163	0,143	0,224	0,528	0,215	0,204	0,222	0,217	0,255	0,226	0,182		0,276	0,27
H.-Gour.	0,264	0,324	0,312	0,363	0,301	0,404	0,245	0,277	0,291	0,353	0,545	0,29	0,234	0,229	0,2	0,245	0,288	0,301	0,276		0,101
B.-Gour.	0,255	0,309	0,304	0,343	0,305	0,391	0,248	0,271	0,283	0,337	0,537	0,288	0,241	0,235	0,192	0,26	0,294	0,323	0,27	0,101	

Fig. 4: Results following the application of Levenshtein's algorithm

It may be striking that most numbers are very close to zero. This means that Georgian dialects exhibit a closer proximity to one another than is typically observed among dialects in other languages. A further analysis shows that the differences between the coefficients corroborate the conclusions of the Georgian dialectologists: there is a clear split between Eastern and Western dialects, as was convincingly argued by Jorbenadze (1989). The diagram in Fig. 5 shows the corresponding hierarchical clustering dendogram.

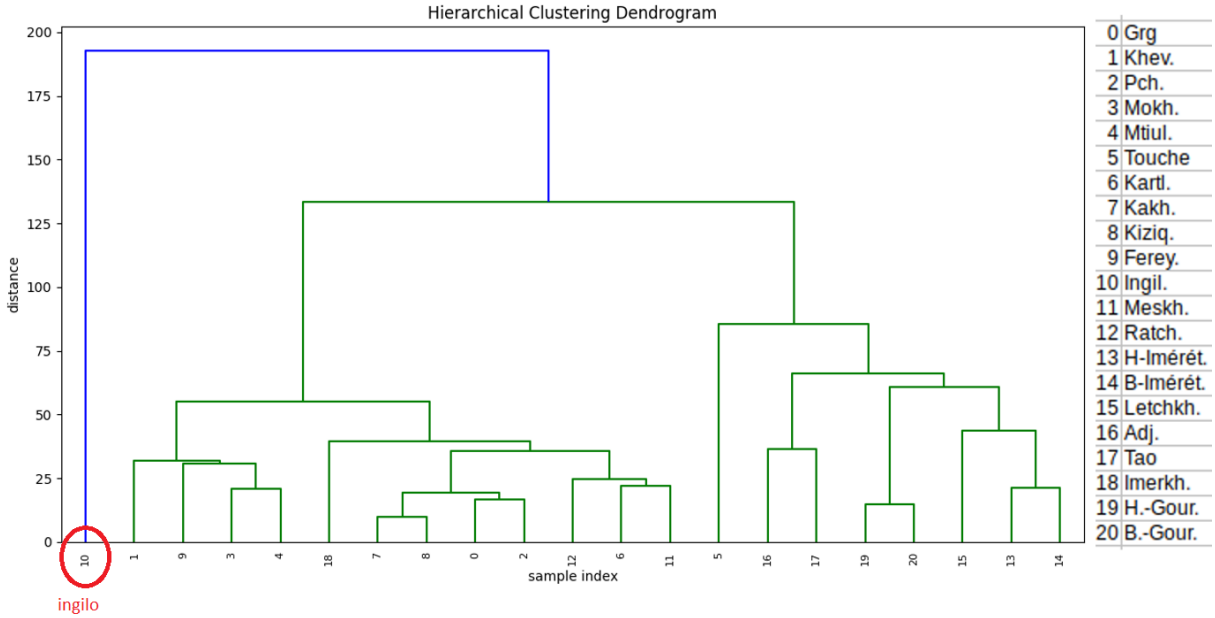

Fig. 5: Results with hierarchical clustering dendogram

These conclusions can also be presented differently, namely by applying ponderation criteria, as shown in Fig. 6.

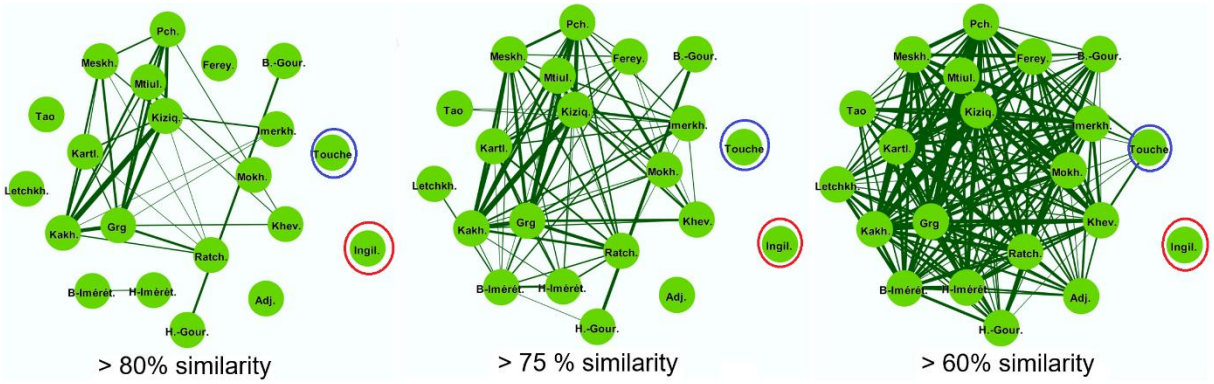

Fig. 6: Results after application of ponderation criteria

The steps of the analysis and processing are described in detail in Léonard (2019). It suffices to mention here that the most salient point is the peripheral position of Ingilo (indicated by a red circle in Fig. 5), i.e. the Georgian dialect spoken in Azerbaijan, and Tush (indicated in purple), an Eastern Mountain dialect. Interestingly, Fereydani, the dialect spoken in Iran and therefore the only dialect that has had no direct contact with Standard Georgian for three centuries, does not exhibit such a divergence. It may be crucial that Tush is in contact with other Mountain dialects, prompting the need to determine what makes it distinct.

This is what the application of Levenshtein Distances can contribute to the study of Georgian dialects. However, it is important to note that this algorithm has its limitations (in particular, it places emphasis on the initial part of words, whereas in Kartvelian, the final part is very important). It would be interesting to explore the data using other distance algorithms, such as the Jaro-Winkler distance.[9] The lexical base that I created can easily be used for other types of distance measurements.

3.3 Diachronic Studies

The Database is also designed to be used for diachronic purposes. Dialects are often underestimated in etymological works, compared to other languages of the same family (which also have their dialects). By taking into account dialectal data and leveraging the abundance of forms it offers, the Database is intended to provide new materials that can refine research in this field.

The diagram displayed in Fig. 7 is another extract from the Database, illustrating what could be a comparative lexicon table. The goal is to establish regular sound correspondences and reconstruct proto-forms.

Etym	Grg	Old	Khev	Psh	Tia	Mok	Mtiu	Guc	Tus	Kar	Kakh.	Kiziq.	Fer	Ing.	Djav	Samt	Ratch	Im	Lech.	Adj.	Tao	Imer	Gur.	Ming	Laz	Svan	Meanin
*ḳac	ḳac	ḳaci	ḳaci	ḳac	ḳac	ḳaci	ḳaci	ḳac	ḳaci	ḳaci	ḳaci	ḳaci	ḳaci	ḳac	ḳaci	ḳaci	ḳaci	ḳac	ḳaci	ḳaci	ḳaci	ḳaci	ḳaci	ḳoči	ḳoč	čäši,	man
*ḳ̇lde	ḳlde	ḳlde	ḳilde	ḳlde		ḳlde	ḳlde		ḳdej	ḳil	ḳlde	ḳlde	[daamana		ḳlde	ḳlde	ḳde,	ḳd	ḳṭe, ḳ	ḳild	ḳild	ḳlde	ḳṭei	ḳyrde,	ḳi	kylde	rock
*ḳbil	ḳbil	ḳbili	ḳbili	ḳbili			ḳvil	ḳvil	ḳmi	ḳbil	ḳbili	ḳbili	ḳbili			ḳbili		ḳbi	ḳṗili,	ḳybi	ḳybi	ḳbili		ḳəbir	ḳibi	(r), ḳi	tooth
*man	man	man	mam	mama		man	man	mama		mar	mama	mama	man	mamaj			mam	ma	mama	[baba, babo]		man		mum	[mu	mū	father

Fig. 7: Sample comparative lexicon table extracted from the database

4. Prospective research

The work accomplished so far represents merely the first step in a much larger undertaking. This mission must be continued and made available to as many researchers as possible. To achieve this, the foremost priority is, of course, enriching the database by filling in the empty boxes. This tedious but indispensable task can be carried out according to the methodology which was presented in section 2.3 but other methods are also worth considering.

At the same time, there is potential benefit in incorporating dialectological data from other Kartvelian languages. Applying a similar dialectal subdivision for Megrelian, Laz and Svan, as done for Georgian, would be advantageous, as well as easier since on the one hand the dialects are fewer in number, and on the other hand their classification is better established.

[9] I thank Gabriel Képéklian for providing insights into these perspectives.

Furthermore, with regard to fieldwork, elicitations are more manageable than for Georgian dialects, as the units differ more, and the risk of interference with Standard Georgian is smaller.

Once a significant number of items have been gathered (without waiting for the database to be exhaustively filled), the next step is to archive the data online. Here, a digital approach is highly suitable because the large number of cells (especially columns, due to the large number of dialect varieties) makes paper printing practically impossible. Another advantage of hosting the database online would the possibility to open it for continual enriching over time. In any case, the archive must include a reference section listing all the sources used for each variety.

5. Conclusion

The current paper presents work in progress focused on creating a lexical database of Georgian cognates that I intend to make available to researchers with the aim of promoting the integration of dialectal data in Kartvelological studies. After discussing the main issues and methodological aspects of data collecting and formatting, I have provided several examples of application, among them of Levenshtein Distances. Other potential uses could entail using other distance algorithms or pursuing diachronic comparison. In the future, my aim is to make this base as comprehensive as possible and archive it online. The long-term goal is to create an open database to be gradually expanded.

References

Beijering, Gooskens & Heeringa (2008): Karin B., Charlotte G., Wilbert H., "Predicting intelligibility and perceived linguistic distance by means of the Levenshtein algorithm", *Linguistics in the Netherlands* 25, 13–24.

Beridze & Nadaraia (2009): Marine B., David N., "The corpus of Georgian dialects", in *Proceedings of the NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference*, Bratislava: Tribun, 25–35.

Beridze et al. (2018): Marine B., Zakharia Pourtskhvanidze, Lia Bakuradze, David Nadaraia, "Interactive Visualization of Dialectal Lexis Perspective of Research Using the Example of Georgian Electronic Dialect Atlas", in *Proceedings of the XVIII EURALEX International Congress*, Ljubljana: Ljubljana University Press, Faculty of Arts, 931–939.

Chikobava (1952): არნოლდ ჩიქობავა, ენათმეცნიერების შესავალი [Introduction to Linguistics], Tbilisi: Tbilisi University Press. https://iverieli.nplg.gov.ge/handle/1234/411129.

Dzidziguri (1970): შოთა ძიძიგური, ქართული დიალექტოლოგიური ძიებანი [Georgian Dialectological Research]. Tbilisi: Ganatleba.

Gigineishvili, Topuria & Kavtaradze (1961): ივანე გიგინეიშვილი, ვარლამ თოფურია, ივანე ქავთარაძე, ქართული დიალექტოლოგია [Georgian Dialectology]. Tbilisi: Tbilisi University Press. https://iverieli.nplg.gov.ge/handle/1234/506997.

Jorbenadze (1989): ბესარიონ ჯორბენაძე, ქართული დიალექტოლოგია [Georgian Dialectology], I. Tbilisi: Mecniereba.

Léonard (2019a): Jean-Léo L., "Méthodes pour l'analyse et la documentation des langues et dialectes kartvèles, à la lumière de la Théorie des Dynamiques Langagières / Théorie de la Complexité [Methods for the Analysis and Documentation of Kartvelian Languages and Dialects, in Light

of the Theory of Language Dynamics / Complexity Theory]", in *Proceedings of the Fifth International Scientific Conference Language and Culture*. Kutaisi: Akaki Tsereteli State University, 625–650.

— (2019b): "Le projet LaDyCa (Language Dynamics in the Caucasus, IDEX Emergence 2017–18): avancées et résultats", *Kadmos* 9, 258–284. https://doi.org/10.32859/kadmos/9/258-284.

Léonard & Makharoblidze (2022): Jean-Léo L. & Tamar M., "Hints at Georgian Dialect History: A Study in Miniature", in: Tamar Makharoblidze (ed.), *Issues in Kartvelian Studies*, Wilmington: Vernon Press, 3–28.

Mart'irosovi (1985): არამ მარტიროსოვი, "ქართული დიალექტური ლექსიკის შესწავლისა და ლექსიკონების შედგენის ძირითადი საკითხები" [Key Issues in Studying Georgian Dialect Vocabulary and Compiling Dictionaries]. *Ibero-Caucasian Linguistics* 23, 139–148.

Picard, Gérardin & Léonard (2018): Flore P., Hélène G., Jean-Léo L., "Levenshtein Algorithm applied to Kartvelian Phonological Data: A Report on the LaDyCa Project (Language Dynamics in the Caucasus, IDEX Emergence, 2017–18): Methods in Theoretical and Empirical Dialectology", paper presented at *Methods for Endangered Kartvelian Varieties Documentation & Analysis according to Language Dynamics Theory* II, 23 February 2018,. Tbilisi: Ilia State University.

Shanidze (1957): Акаки Шанидзе, "Принципы классификации грузинских диалектов [Principles of the Classification of Georgian Dialects]", *Proceedings of the Joint Scientific Session of the Academy of Sciences of the Transcaucasian Republics on Social Sciences*, Baku, 828–832.

Online resources

Georgian Dialect Corpus (GDC): http://corpora.co
Georgian National Corpus (GNC): http://gnc.gov.ge
TITUS Database: https://titus.uni-frankfurt.de

დიალექტოლოგიური ლექსიკოლოგიური მონაცემთა ბაზის აგება ქართული ენის ზიარი კოგნატების ანალიზისათვის

ელენე ჟერარდინი (პარიზი)

DOI: https://doi.org/10.62235/dk.3.2024.8511
helene.gerardin@inalco.fr || ORCID: 0009-0003-7406-3901

წინამდებარე ნაშრომში განხილულია ის პრობლემები და გამოწვევები, რომელიც ჩვენ მიერ განხორციელებული პროექტის - **ქართული დიალექტების ლექსიკურ მონაცემთა ბაზის** შექმნას ახლდა თან. ეს გამოწვევები, უპირველეს ყოვლისა, დაკავშირებული იყო მონაცემთა შეგროვებისა და ფორმატირების მეთოდოლოგიური მიდგომების შემუშავებასთან, რომელიც დაწვრილებით არის განხილული სტატიაში.

პროექტის იდეა გაჩნდა IDEX-ის ერთ-ერთი პროექტის Linguistic Dynamics of the Caucasus (LaDyCa) ფარგლებში, რომელიც 2017-2018 წლებში განხორციელდა სორბონის უნივერსიტეტისა (პარიზი) და ილიას სახელმწიფო უნივერსიტეტის (თბილისი) თანამშრომლობით. თავად IDEX-ი („Excellence Initiatives“, ფრანგულად „Initiatives d' Excellence“) წარმოადგენს საფრანგეთის სახელმწიფოს მიერ შექმნილი პროგრამის „ინვესტიციები მომავლისთვის“ (“Investments for the Future”) ნაწილს, რომელიც მიზნად ისახავს კონკურენტუნარიანი მულტიდისციპლინური უმაღლესი საგანმანათლებლო და კვლევითი ინსტიტუტების შექმნას ქვეყანაში.

ქართული ენა წარმოდგენილია ფართო ენობრივი და გეოგრაფიული ვარიანტების სახით, რაც საფუძვლად უდევს ქართული დიალექტების მრავალფეროვნებას. სხვადასხვა ავტორის მიხედვით (იხ. მაგალითად Chikobava 1952, Shanidze 1957, Dzidziguri 1970 და Jorbenadze 1989), ქართულში გამოიყოფა 15–დან 20-მდე დიალექტი ან ქვედიალექტი, რომელიც კლასიფიკაციის განსხვავებულ პრინციპებს ეფუძნება.

პროექტი მიზნად ისახავდა ქართულ დიალექტებში დადასტურებული საერთო ლექსიკური მნიშვნელობის მქონე სიტყვათა შეგროვებასა და ლექსიკურ მონაცემთა ბაზის შექმნას ღია რესურსის სახით, რომელიც გამოსადეგი იქნებოდა არა მარტო დიალექტოლოგიური მასალის არქივირებისა და შემდგომი განვითარებისათვის, არამედ ქართულის ენობრივი ვარიაციების შესასწავლად, ასევე ქართული დიალექტების კლასიფიკაციის დასახვეწად და ენის დიაქრონიის შესწავლისა და რეკონსტრუქციის გასაადვილებლად. გარდა ამისა, ქართული დიალექტების ონლაინარქივირება შემდგომში მისი გაფართოვებისა და დახვეწის შესაძლებლობასაც იძლევა.

ქართული დიალექტების ლექსიკურ მონაცემთა ბაზის ბირთვს წარმოადგენს ბაზის სტრუქტურა, რომელიც აერთიანებს ქართული დიალექტების საერთო ლექსიკური მნიშვნელობის მქონე სიტყვებს (საზიარო კოგნატებს). ლექსიკური ერთეულების ბაზაში გაერთიანებისას მონაცემები, ერთი მხრივ, გაიფილტრა ნასესხები სიტყვებისაგან, რაც განსაკუთრებით უხვად დასტურდება სხვა

ენებთან კონტაქტში მყოფი დიალექტების შემთხვევაში; მეორე მხრივ, დადგინდა ლექსიკური ერთეულების წარმომადგენლობითი სია, არანაკლებ ასი ელემენტის მოცულობისა.

იმ სირთულეების გათვალისწინებით, რომელიც გაჩნდა ლექსიკური ერთეულების შერჩევის პროცესში, ასევე დიალექტოლოგიური რესურსების სიმცირიდან გამომდინარე, ჩვენ მიერ შექმნილ ბაზაში თავი მოვუყარეთ რესურსებს როგორც ზეპირმეტყველების, ისე ბეჭდური წყაროებიდან. უნდა აღინიშნოს, რომ ბეჭდური წყაროები ძირითადად მოიცავს ლექსიკონებსა და სიტყვების სიებს, რომლებიც ჩვეულებრივ თან ერთვის ტექსტების კრებულებს. ასეთი სიების შემთხვევაში პრობლემას გვიქმნიდა ის გარემოება, რომ სიაში ძირითადად შესულია ისეთი ლექსემები, რომლებიც ყველაზე მეტად არიან დაშორებული სტანდარტს, მაშინ როდესაც საზიარო ლექსიკური ერთეულები ხშირად საერთოდ არ არის ასახული ლექსემათა ჩამონათვალში. გარდა ამისა, მონაცემებს ვავსებდით გამოქვეყნებულ ტექსტებსა და ონლაინკორპუსებში დაცული მასალებით.

მონაცემების შეგროვების შემდეგ რესურსები გადავიდა მონაცემთა ბაზაში და განთავსდა ცხრილებში, სადაც ლექსიკური ერთეულები კლასიფიცირებულია დიალექტის მიხედვით. თითოეულ ერთეულს თან ახლავს თარგმანი ინგლისურ და ფრანგულ ენებზე. ლექსიკური ერთეულები წარმოდგენილია სტანდარტული ქართული ლემის მიხედვით და დალაგებულია ანბანური თანმიმდევრობით. შედარებისა და დიაქრონიული ანალიზის გასაადვილებლად, ლექსიკურ ერთეულებს ჩვენ ასევე დავურთეთ შესაბამისი ფორმები ძველ ქართულში და რეკონსტრუირებული ქართველური ეტიმონები (ასეთის არსებობის შემთხვევაში). გარდა ამისა, ლექსიკურ მონაცემთა ბაზის გასაფართოვებლად მასალას დაემატა სხვა ქართველური ენების - მეგრულის, ლაზურისა და სვანურის ლექსიკური ერთეულები, რომლებიც ცხრილში განთავსდა ქართული დიალექტების მარჯვნივ განლაგებულ სვეტებში.

მონაცემთა ბაზის მიმდინარე ვერსიის სტრუქტურის აღწერის შემდეგ სტატიაში განხილულია მისი გამოყენების შესაძლებლობები და უპირატესობა ლინგვისტური კვლევებისათვის. სტატიაში წარმოდგენილია ლექსიკურ მონაცემთა ბაზაში ლევენშტაინის დისტანციების გამოთვლების შედეგად მიღებული კვლევის შედეგები და დიალექტების დიაქრონიული და დიატოპიური შედარება.

ლევენშტაინის ალგორითმის შედეგად მიღებული სტატისტიკური შედეგები, რომელიც პროექტის წევრის, ლაურა პიკარდის მიერ იქნა გამოანგარიშებული, ცალსახად გვიჩვენებს, რომ ქართული დიალექტები უფრო ახლოს არიან ერთმანეთთან, ვიდრე ეს ჩვეულებრივ შეინიშნება სხვა ენათა დიალექტების შედარებისას. კოეფიციენტებს შორის განსხვავება კი ემპირიულად ადასტურებს ქართველი დიალექტოლოგების დასკვნებს, რომ აღმოსავლურ და დასავლურ დიალექტებს შორის აშკარაა არსობრივი სხვაობა (იხ. დიაგრამა 5-ზე წარმოდგენილი იერარქიული კლასტერული დენდოგრამა).

აქ აუცილებლად უნდა აღინიშნოს ორი დიალექტის პერიფერიული პოზიციონირების ფაქტი; კერძოდ, ყველაზე გამოკვეთილი სხვაობა დასტურდება აზერბაიჯანში გავრცელებულ ინგილოურ დიალექტსა და აღმოსავლური მთის

დიალექტ თუშურში, მაშინ როდესაც ირანში გავრცელებული ფერეიდნული დიალექტი (ერთადერთი დიალექტი, რომელსაც სამი საუკუნის განმავლობაში არ ჰქონია შეხება სტანდარტულ ქართულთან), არ ავლენს ასეთ განსხვავებას. შეიძლება აქ გადამწყვეტი ფაქტორი იყოს თუშური დიალექტის კონტაქტი სხვა მთის დიალექტებთან, რაც მოითხოვს დამატებით კვლევას, თუ რა განასხვავებს მთის დიალექტებს ერთმანეთისაგან.

სტატიაში წარმოდგენილი დასკვნები ეყრდნობა ლევენშტაინის დისტანციების გაზომვის მეთოდს. თუმცა, უნდა აღვნიშნოთ, რომ ლევენშტაინის ალგორითმს გარკვეული ხარვეზები გააჩნია (იგი ფოკუსირებულია სიტყვის საწყის სეგმენტზე, მაშინ როდესაც ქართველური ენებისათვის გაცილებით მნიშვნელოვანია სიტყვის ბოლო ნაწილი). ამიტომ საინტერესო იქნებოდა მასალის ანალიზი დისტაციების სხვა ალგორითმის, მაგალითად იარო ვინქლერის დისტანციების ალგორითმის გამოყენებით. მითუმეტეს, რომ ჩვენ მიერ შექმნილი ქართული დიალექტების ლექსიკურ მონაცემთა ბაზა შეიძლება ადვილად იქნეს გამოყენებული სხვა ტიპის დისტანციების გამომთვლელი ალგორითმებისათვის.

სამომავლოდ ჩემი მიზანია შევავსო ქართული დიალექტების ლექსიკურ მონაცემთა ბაზა და შევქმნა ონლაინარქივი. რაც შეეხება გრძელვადიან მიზნებს, **ქართული დიალექტების ლექსიკურ მონაცემთა ბაზის** საფუძველზე უნდა შეიქმნას ღია წვდომის რესურსი, რომელიც შემდგომში თანდათან შეივსება და გაფართოვდება.