Jean Léo Léonard (UPV, Dipralang, Fr)

Hints at Georgian Dialect History: A Study in Miniature

Abstract:
This study proposes a set of hypotheses on the formation of the Georgian Dialect Network (GDN) from the standpoint of *Language Dynamics* and *Complex Systems Theory*, through a *model* articulated on ten fundamental notions from General Dialectology: (L1: *L* standing for *Layer*) *The Dialect Split Layer* (DSL), (L2) *The Buffer Zone Effect* (BZE); (L3) *Variable Bleeding* (VB), i.e. *Relative Chronology* (Scalar Change); (L4) *The Feature Pool Effect*; (L5) *Emerging Isolates* (EI) or singleton or dendrographic outliers; (L6) *The Centre-Periphery Effect* (CPE), i.e. Bartolian centre/periphery interplay; (L7) *Phonolexical Endemic Patterns* (PLEP); (L8) *Word Geography* (WG), i.e. lexical diffusion; (L9) *Local Semantic Shifts* (LSS); (L10) *External Factors* (EF). A database of 243 cognates compiled from classical sources referenced in Georgian dialectology enabled quantitative tests for dialect clustering using *Gabmap* (i.e. Levenshtein algorithm), with particular attention paid to *Hierarchical Clustering, Difference Maps, Multiscalar Dimension Plots*, and *Weighted Average + Group Average Probability* clouds). The results from this first endeavour in Dialect Dynamics applied to the GDN have provided a certain number of orientations for future research in Georgian dialectology and sociophonetics, in particular on the properties of types L1-4 and L5-6 of this diasystemic topology, which highlight some deep organizational patterns. These leads could be heuristic, and help to lay the groundwork for dialectometry applied to the GDN, in addition to current projects, such as the *Georgian Dialect Corpus*. Further research should focus e.g. on L7-10, in order to explore more superficial levels of diffusional trends.

1. Introduction

Key words: *Language Dynamics, Levenshtein algorithm, Computational Dialectology, Dialectology, Georgian, Georgian Dialect Corpus.*

Anyone familiar with the history of dialectology will have noticed that the title of this chapter is a tribute to Sarah Gudschinsky's seminal paper (1958) on the history of a Mesoamerican language - Mazatec. Moreover, I am further indebted to another impressive piece of scholarship on a similar topic, in this case applied to the Caucasus and a region located in north-eastern Georgia: Johanna Nichols' article (2004) on "The Origin of the Chechen and Ingush: A Study in Alpine Linguistic and Ethnic Geography". These contributions accomplished a decisive step forward in their respective fields: on the one hand, designing an elegant model for the description of geolinguistic dynamics stemming from both internal and external factors, on the other hand, applying Victor

Murra's concept of the Vertical Archipelago (Murra 1956, 1985) to the Caucasus, making it possible to encompass a vast amount of phenomena pertaining to population dynamics, settlements and interactions in space and time within a unified ecological framework. A third source of inspiration also triggered the tentative approach I present here, on Georgian dialect history and dynamics: Alexei Kassian's endeavour (2015) to test algorithmic complexity on Lezgian languages (North Caucasus). Kassian tests a wide array of quantitative methods (distance-based, such as StralingNJ, NJ, UPGMA, versus character-based, such as Bayesian MCMC, UPM) to match the available "standard classification" or "received taxonomy" of Lezgian languages, obtained through the qualitative analysis of cognates (in other terms through isoglosses, i.e. types and trivial characters). Kassian's approach is reminiscent of the Popperian falsificationist methodology, which considers that scientific knowledge progresses through plausible hypotheses and results which are then confirmed or denied by additional evidence or alternative methods (Popper 1934, 1963). The results of any scientific inquiry must not be considered as set in stone but must be designed so as to yield a fruitful response to validity testing, and to falsificatory procedures (here, in sections 3.1-3, confronting the output of Figure 1 in section 2).

I will refer to *Guschinsky's Model of Dialect Dynamics*[1] as GMDD, the premises of which are sketched out in (1). In terms of general systemics, L1-3 (L stands for *Layer*) are ascending variables, i.e. competing to enhance the dialect network inner diversity, and providing its external shape (*emergence*). L4 and L7 entail *flows of information* and models (structural patterns, paradigms). L5 is typically a generative parameter, either local or regional. L6 can be defined as a strongly dynamic parameter of self-organization (*autopoesis*), which modifies the spatial structure and the thread of the dialect network - as an anamorphosis distorting physical space, to give shape to the topology of the network. L8 is typically extensive and diffusional (*centrifugal flow of informatio*n), whereas L9 is typically intensive and structurally introverted (*centripetal self-organization*). L10 should be considered as a model of more or less compelling external pressure (political power and the conditioning of social agentivity and interactivity between speakers).

(1) The GMDD applied to the Georgian Dialect Network (GDN): L = (geolinguistic or areal) Layer.

L1: The Dialect Split Layer (DSL). This level of analysis accounts for the main divisions of a dialect network, such as West versus Centre & East in the "Standard Classification" of Georgian dialects (SCGD). This major level of division involves e.g. sound changes akin to the "Neogrammarian Laws",[2] such as the famous hissing-hushing obstruent division[3] or the /*a*/ vs. /*o*/ opposition between Georgian and Zan at Kartvelian level, but also within the Georgian Dialect Network (GDN). The geolinguistic contrast between varieties preserving modal and glottalized uvulars, such as /*q, q̇*/ also works to some extent as a L1 variable, although allophonic variation is fairly frequent in ordinary speech and may vary depending on idiolects and sources, resulting in some subsequent blurring from a synchronic standpoint.

[1] *Dialect Dynamics* is a component of *Language Dynamics*, understood here as in Heinsalu & *al*. (2020).
[2] See e.g. Machavariani (1965), Gamkrelidze (2005).
[3] Cf. Asatiani (2008).

L2: the Buffer Zone Effect (BZE). Once main dialect divisions appear in a common language as a result of DSL variables, a trend towards areal overlapping generally occurs as a by-product of social interaction, contextual or free variation between neighbouring dialects. At a higher level, the Central area in the GDN stands as a buffer zone between East and West in the domain.

L3: Variable Bleeding (VB). A DSL (i.e. L1, above) entails subsequent complexification or simplification of the sound law or structural change at stake – relative chronology generally accounts for the intricacy and the hierarchisation of this process. This trend may also feed the previous one (BZE), or focus on one or both divisions, enhancing further splits and differentiation, at sub-dialect levels (the north-eastern area in the East and the Gurian area in the South-West provide good examples of this trend).

L4: **The Feature Pool[4] Effect, First Grade (FPE I)**: structural polymorphism, especially systemic trends towards a certain probability of contextual or free variation.

L5: **Emerging Isolates (EI)**, such as the rise of single dialects,[5] out of local rules and idiosyncratic trends (e.g. *bVr* > *brV* metathesis in the South-Western dialect, see *bevri* 'much' –item 1, Table 2 below). The labiopalatalisation of front vowels in Ingilian for items 1 'much', 3 'mill' and 7 'side' in Table 2 (Grg *bevri* INGIL[6] ***bövrü***; Grg *ciskvili* INGIL *c**ü**ck**ü**l*; Grg *gverdi* INGIL *g**ö**rd*) provides further examples, for a singleton variety, through contact with Azeri Turkish, due to typological areal convergence).

L6: **The Centre-Periphery Effect (CPE)**, or the *Bartolian effect* (Bartoli 1945), which predicts the high probability of centrifugal versus centripetal phenomena of diffusion in a given geolinguistic space. According to this trend, central innovations expand in space and time, yet are hindered or blocked by retentions at the periphery. Dialect zones located in compartimentalised highlands, such as the North East PSH, MOX, MTIU/GUD and TUSH, or located in areas far from the main centres of diffusion of leading town dialects, or whose populations have migrated abroad (INGIL, FEREY) tend to belong to these so called 'lateral zones'. The CPE is cyclical: it may work at macrolevel (the whole GDN) or at regional levels (e.g. H&L IMR, and, to some extent, L-GUR behave as centrifugal, leading dialects interacting with surrounding sub-dialects, such as LCHK or ADJA). The main centre of gravity of the GDN indeed lies around the KAR dialect in the central part of the country, in strong historical interaction with the KAXET dialect, in the East.

L7: **Phonolexical Endemic Patterns (PLEP)** could be called *The Feature Pool Effect, second Grade* (FPE II). In this case, the diversity of morphophonological rules of surface realizations, at what can be called the phonolexical level, blurs variation patterns, such as item 8 'come' aorist 3Sg, in Table 2: GRG *movida*; XEV *mavi*; PSH *moida*; MOX *movida*; MTIU *moida*; RACH *mevida*; H-IMR *mevida*; L-IMR *mouda*; IMRX *mojda*; GUR *mevida*. Numerous microscale interactions occur in the syllabic template between adjacent onsets and nuclei, out of homorganic contact of *-ov-* in the stem, with redundant [Labial] feature for *o* and *v*, making geolinguistic patterns undecidable.

[4] A *Feature pool* can be defined as a complex set of variants available for the same function, more or less freely available to speakers of a dialect network, as in a Creole continuum, or mingled together in mixed varieties. Diversity of free variants in competition blurs dialect frontiers, and makes up a *pool* –or a *pond*– of variation quite different from the "traditional dialect/sociolect" settings (see Mufwene 2001, 2013).

[5] The term 'isolate' here points to any dialect variety which strongly differs from average variation within the Georgian Dialect Continuum. We by no means refer to a 'phylogenetic isolate', here.

[6] Abbreviations for Georgian dialect varieties are given in Table 1 below.

Most of these microtrends do not make up a clear-cut regular sound law, resulting in another type of dialectical feature pool.
L8: **Word Geography (WG)**, or lexical diffusion. Items may not be comparable cognates either because of different structural choices in the inherited lexicon and/or semantic shifts (e.g. *testa* versus *capo, cabeza, chef* in Romance languages for 'head'), or out of complex processes of derivation, blurring the calculability of cognates in the word list (at least for automated dialectometry).
L9: **Local Semantic Shifts** (**LSS**): not examined here; nevertheless, some of these items may still be used for comparison at phonological level in our data processing.
L10: **External Factors (EF)**: history, geography and geopolitics, competing native or foreign hegemony and superimposed contact languages, etc. In this respect, one could argue that from the standpoint of Braudelian "Long Duration", the main division between Kartvelian languages (Svan and Zan in the West, Georgian in the Central and Eastern part of contemporary Georgia) brings to mind the geopolitic contrast, in ancient times, between the kingdom of Colchis versus the Kingdom of Iberia in Western and Central-Eastern Georgian –with Argveti as a former "buffer zone". Later on, the same division within the GDN still holds, with strong influence from foreign hegemonies (Turkey in the West, Persia in the East). Although this is a fascinating issue, as the title of this paper suggests: *hints* at Georgian dialect history, I will not go into detail here, but I hope the diversity of interpretations flowing from the various dialectometric topologies I highlight will be useful for acknowledged specialists such as historians and geographers.

In section 2, I will first describe the structure of the data used in this chapter. I will present the canonical classification of Georgian dialects, as a compass to make our way into the maze of the Georgian dialect network. I will provide samples of the data (Tables 2-4), and give an account of the processing of the totality of the database, before testing results on the basis of smaller sets of cognates. I will therefore survey three subsets of data, corresponding to groupings of phonological variables. First, the hissing-hushing fricatives and affricates, as a prototypical DSL (Dialect Split Layer) variable; second, the uvular stops and ejectives, as a false DSL and a genuine polymorphic and unstable variable (FPE); third, the labialized stops (*Cv*), which will provide evidence of the Bartolian dynamic field (CPE), and highlight some trends of the FPE and EI –to a certain extent. I will close the chapter by exploring what General Dialectology may contribute to a specific field of research such as Kartvelian studies, and how, in turn, Kartvelian studies may considerably enrich General Dialectology.

2. Data processing: the standard taxonomy of Georgian dialects vs. alternative topologies

I will use a database compiled by Hélène Gérardin (Inalco, Paris) in 2018, within the framework of the IDEX EMERGENCE project LaDyCa (*Language Dynamics in the Caucasus*), a *Complexity Theory* and *Language Dynamics* project carried out at Sorbonne University in 2017-18 (Léonard 2019, 2017), and two sets of dialectometric results: from Flore Picard's quantitative processing of 243 cognates on *Gabmap* and *Gephi* (Picard & al. 2018). Table 1 shows the SCGD (i.e. the standard dialect network taxonomy of Georgian

dialects) according to Gigineishvili, Topuria, and K'avtaradze (1961).[7] Table 2 provides a sample of the LaDyCa 2018 database by Hélène Gérardin (Inalco). Abbreviations used in this chapter to refer to the 22 dialect varieties surveyed are given in Table 1, within brackets (IMR, GUR, KAR, etc.).

Table 1. The Gigineishvili, Topuria, and K'avtaradze (1961) Georgian dialect classification

West		**East**		
Northwest dialects	**Southwest dialects**	**Central dialects**	**Northeast dialects**	**Eastern dialects**
Imeretian (Imeruli, იმერული) **[IMR]**	**Gurian** (Guruli, გურული) **[GUR]**	**Kartlian** (Kartluri, ქართლური) **[KAR]**	**Mokhevian** (Mokheuri, მოხეური) **[MOX]**	**Kakhetian** (Kakhuri, კახური) **[KAXET]**
Lechkhumian (Lečkhumuri, ლეჩხუმური) **[LCHX]**	Adjarian (Ačaruli, აჭარული) **[ADJA]**	**Meskhian** (Meskhuri, მესხური) **[MESH]**	**Mtiuletian-Gudamaqrian** (Mtiulur-Gudamaqruli, მთიულურ-გუდამაყრული) **[MTIU]** **[GUD]**	**Tianetian** (Tianeturi, თიანეთური)
Rachan (Račuli, რაჭული) **[RACH]**	**Imerkhevian** (Imerkheuli, იმერხეული) [Turkey, Imerkhevi] **[IMRX]**	**Javakhian** (Javakhuri, ჯავახური)	**Khevsurian** (Khevsuruli, ხევსურული) **[XEV]**	**Ingiloan** (Ingilouri, ინგილოური) [NW Azerbaijan] **[INGIL]**
	Taoan[8] **[TAO]**		**Pshavian** (Phšauri, ფშაური) **[PSH]**	**Fereydanian** (Phereidnuli, ფერეიდნული) [Iran] **[FEREY]**
			Tushetian (Tušuri, თუშური) **[TUSH]**	

Source: https://en.wikipedia.org/wiki/Georgian_dialects[9] (modified)

One of the technical problems arising from this kind of second-hand compilation of dialectical cognates from a complex set of written sources (see Appendix 2 for a sample)

[7] NB: also quoted in a reference paper, Tuite (1989:5).
[8] Not mentioned in the source of Table 1, but documented in our database. See Chokharadze & *al*. 2018 on the social history of this south-western variety, mostly spoken in Eastern Turkey.
[9] See map Online at https://commons.wikimedia.org/wiki/File:Georgian_dialects.svg.

lies in the empty cells (indicated by a bar ___ in the table). A first round of data processing was therefore carried out using a cascade model, separately comparing dyads of each variety to the whole corpus, in order to then unify the results within a single matrix (see Appendix 1).

Table 2. A sample of the LaDyCa database, compiled by Hélène Gérardin (Inalco)

	1	2	3	4	5	6	7	8
	'much'	**'we'**	**'mill'**	**'horse'**	**'one'**	**'wolf'**	**'side'**	**'come'** AOR3Sg
GRG	*bevri*	*čven*	*ṣiskvili*	*cxeni*	*erti*	*mgeli*	*gverdi*	*movida*
XEV	*bevr*	*čven*	*ṣiskvili*	*cxeni*	*erti*	*mgeli*	*gverdi*	*mavi*
PSH	*bevri*	*čven*	*ṣiskvili*	*cxeni*	*erti*	*mgeli*	*gverdi*	*moida*
MOX	*bevri*	*čon*	*ṣiskwili*	*sxeni*	*erti*		*gördi*	*movida*
MTIU	*bevri*	*čon*	*ṣviskvili*	*cxeni*	*erti*	*mgeli*	*gordi*	*moida*
GUD	___	*čven*	*ṣviskvili*	*cxeni*	*erti*	___	___	___
TUSH	___	___	___	*cxenĭ*	*ertĭ*	*geli*	*gverdĭ*	*moid*
KAR	*bevri*	*čwen*	*ṣiskvili*	*cxeni*	*erti*	*geli*	*gwerdi*	*moida*
KAXET	*bewri*	*čwen*	*ṣiskvili*	*cxeni*	*erti*	*mgeli*	*gverdi*	*moida*
KIZIQ	*bewri*	*čwen*	*ṣiskvili*	*cxeni*	*erti*	*mgeli*	*gverdi*	*moida*
FEREY	*bevri*	*čön*	*ṣiskili*	*cxeni*	*erti*	*geli*	*gerdi*	*moida*
INGIL	*bövrü*	*čon*	*cückül*	*cxen*	*er*	*gel*	*görd*	___
MESH	*bevri*	*čön*	*ṣiksvili*	*cxeni*	*jerti*	*ngeli*	*gördi*	___
RACH	*bevri*	*čven*	*ṣiskvili*	*cxeni*	*erti*	*geli*	*gverdi*	*mevida*
H-IMR	*brevi*	*čwen*	*ṣiskvili*	*cxeni*	*jerti*	*geli*	*gverdi*	*mevida*
L-IMR	*brevi*	*čwen*	*ṣisḳvili*	*cxeni*	*jerti*	*geli*	*gverdi*	*mouda*
LCHX	*brevli*	*čvene*	___	___	___	*geli*	*gverdi*	___
ADJA	*bewri*	*čven*	*cxeli*	*cxvari*	*ʔēti*	___	___	___
TAO	*bewri*	*čwen*	*ṣisḳwili*	___	*ēti*	*geli*	___	___
IMRX	*bewri*	*čwen*	*ṣisḳwili*	*cxeni*	___	*geli*	*gwerdi*	*mojda*
H-GUR	*bewri*	*čwen*	*ṣiskpili*	*cxeni*	*eti*	*geli*	*gverdi*	*mevida*
L-GUR	*breuli*	*čwen*	*ṣiskpili*	*cxeni*	*eti*	*geli*	*gverdi*	*mevida*

The LaDyCa Georgian Dialect Database provides the following set of results (figure 1), for the whole set of cognates (*n* = 243). Beforehand, let me suggest some basic tenets for reading the Gabmap[10] figures, such as in Kassian (op. cit.): first, distinguish between a core-area or a core-cluster and its outlier(s); second: cyclically apply this technique to the inner structure of each set of clades; third, point at a putative Gudschinsky Model (the

[10] Gabmap is free dialectometry software which uses the Levenshtein algorithm, or "editing distance", (Levenshtein 1966) hosted by CLARIN (see http://portal.clarin.nl). See Leinonen & *al*. (2016), accessible at https://www.sciencedirect.com/science/article/pii/S0024384115000315.

GMDD), as a tentative orientation for further potential research, especially from the standpoint of EF (External Factors: L10) -although I will remain cautious in this respect. For instance, the dendrogram in Fig. 1a is based on Ward's Method (hierarchical clustering). I deliberately decided to generate five clades at each step of the computation, no more, in order to match the standard taxonomy for the GDN, as exposed in Table 1.

Figure 1. The LaDyCa Georgian Dialect Database: 243 cognates

Figure 1a. Ward's Method (243 items)

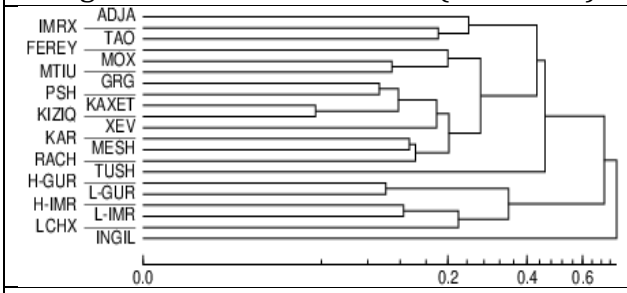


Figure 1b. Statistical distribution of differences

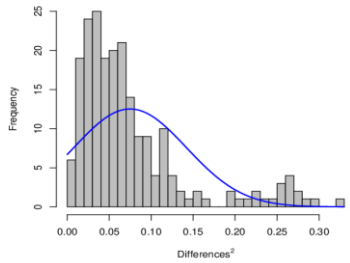


Figure 1c. Multidimensional scaling (r = 0.88)

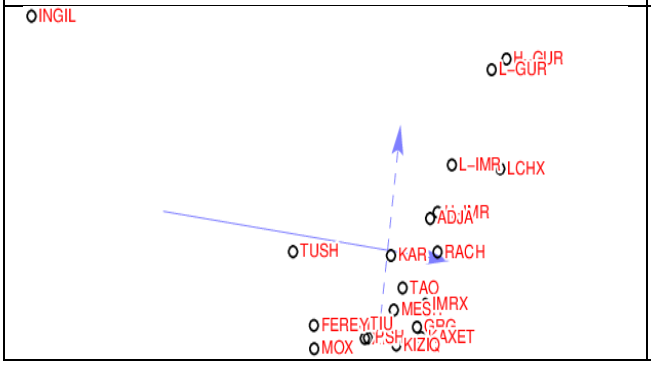
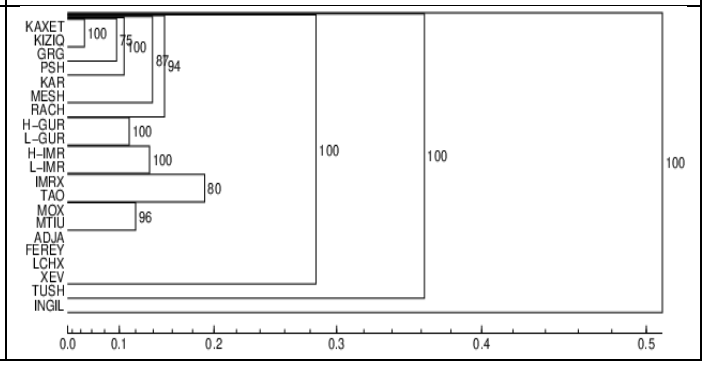


Figure 1d. Probabilistic dendrogram (Weighed average + Group Average; limit 60%, noise: 0.2)



At first sight, there is a massive core in the upper part of the tree, from FEREY to RACH, including most Eastern and Central varieties of the GDN. Three peripheral south-western varieties make up its outlier, which can be explained as an obvious expression of Bartolian CPE. The whole complex has TUSH, a typical isolate (EI) belonging to the eastern complex, as a peripheral outlier. Now, the core of western Georgian appears clearly as a competing centre of innovations, with GUR on the one hand, and IMR on the other hand. LCHX happens to play the same role as an outlier to this group as TUSH does for the core-cluster of Central-Eastern Georgian (as a kernel group), associated to the distant south-western periphery. Now, if we apply GMDD -our geolinguistic modelling of our Gabmap testing output- to each set of clades, the central-eastern core shows intricate and massive topology, which suggests this area has been evolving toward a Feature Pool (L4: FPE I), rather than a confederation of clearly distinct dialects -whereas the Western segment of the GDN seems to have split into a twofold network: on the one hand, a recessive, peripheral zone, associated to the central-eastern complex, on the other hand, a fairly innovative, consistent and compact central-western confederation of dialects, including IMR/LCHK and GUR (cf. L6: CPE). The remainder, INGIL, at the periphery of the whole GDN, should be considered as a singleton, embedded in a strong field of interference from Azeri abroad. Both core areas emerge as competitors, with satellite isolates as outliers of their clades: respectively TUSH (which has undergone intense contact with Tova-Tush, a

Nakh-Daghestanian language) and INGIL (intense contact with a Turkish language: Azeri) ‑here, language contact as an EF (external Factor) goes without saying.

The subsequent figures 1b-d provide additional information about the statistical structure of the sample (Fig. 1b) and the probability of the inner structure of the topology (Fig. 1d). In short, the former shows that the sample is strongly negatively skewed (tail to the left: smaller or rather simplex differences occur more frequently than complex, heterogeneous differences); the latter reveals interesting patterns, such as the IMRX/TAO cluster in the centre of the diagram, which more optimally matches the standard taxonomy (i.e. SCGD) than in the Figure 1 dendrogram. Figure 1c strikingly confirms the trend of the bulk of central-eastern dialects to make up a Feature Pool (L6: Feature Pool I) rather than a clear-cut segment of the dialect network with competing components. The lower arrow points to this trend, while the upper one confirms that the GDN unfolds as a dialect network proper (instead of a Feature Pool) as one proceeds to the West: first to ADJA, then to IMR, and ultimately to GUR. The position of INGIL at the opposite side of the scale, to the left, confirms its status as an emerging isolate (EI) ‑through exogenous contact, as already mentioned. Fig. 1d has the advantage of summing up two alternative computing techniques (Group Average and Weighted Average), making discrete blocks more salient within the inner dendrographic structure. I will therefore systematically use the following four data representations: hierarchical clustering, statistical distribution of differences, MDS plotting and probabilistic dendrogram.

In other words, what this set of data shows does not basically contradict the standard taxonomy, as it opposes the Western part and the Central-Eastern part of the GDN. Nevertheless, it blurs the inner frontiers and subdivisions, and it splits the Western area in two. It therefore fulfils its promises: quantitative methods are better designed to enrich and challenge qualitative methods than to match them straightforwardly. One expects heuristic values ‑*hints*‑ from their results, instead of a faithful picture of what was already available public knowledge and understanding. The GMDD provides inspiring insights into diasystemic trends evolving or having evolved at different stages of the history of the GDN, from the initial putative DSL (the major dialect split) to the Bartolian CPE, through the emergence of competing dialects (IMR/IMRX and GUR) and emerging isolates (TUSH, INGIL).

3. **Testing non-intersecting subsets of data**

Another important consequence of the GMDD method lays in its heuristic properties: the ten layers (L1-10) may be used as a compass to select subsets of variables, and check to what extent they match conventional taxonomies. The product of this confrontation always turns out to be heuristic.

3.1. *Hissing-Hushing obstruents*

Table 3 shows six out of the eight cognates processed for results in Figure 2a-d (to which items 2 ‘we’ and 3 ‘mill’ are added, from Table 2 above).[11] As the diagrams in Figure 2 suggest, with *Gabmap*, small samples can provide fairly relevant and

[11] Nevertheless, even with such a small sample, the data overview yields the following proportions: 22 varieties/objects, 176 instances, 925 characters (including 29 unique characters), 886 tokens (including 30 unique tokens). With any automated endeavour, even small is big…

encouraging results. Nevertheless, fricatives and affricates are not the only variables here: they mingle with other characters, such as vowel tension (lax *ĭ* in TUSH), with a palatal expansion *-j* (approximant) or *-i* (vocoid) of the nucleus for items 13 'brother' GRG ***ʒ****ma* vs. TUSH ***ʒ****maj* and 14 'other' with velarized fricatives: GRG ***sx****va* vs. TUSH ***cx****vaj*, L-IMR ***sx****wai* and GUR ***sx****wai*, endemically. But at this stage of the sampling, I deem necessary to delve into this fine grained variation.

Table 3. Hissing-Hushing obstruents sample from the LaDyCa database

	9	**10**	**11**	**12**	**13**	**14**
	'shadow'	'back'	'straight'	'cross'	'brother'	'other'
GRG	*črdili*	*zurgi*	*sc̣ori*	*ǯvari*	*ʒma*	*sxva*
XEV	*čdili*	*zurgi*	*sc̣ori*	*ǯvari*	*ʒma*	*sxo*
PSH	___	___	*sc̣ori*	*ǯvari*	*ʒma*	*sxva*
MOX	___	*zurgi*	*sc̣ori*	*ǯori*	*ʒma*	*cxo*
MTIU	*čirdili*	*zurgi*	*sc̣ori*	*ǯori*	*ʒma*	*sxo*
GUD	*čirdili*	___	___	*ǯori*	___	*sxo*
TUSH	*čdilĭ*	*zurgĭ*	*sc̣orĭ*	*ǯvarĭ*	*ʒmaj*	*cxvaj*
KAR	*čdili*	*zurgi*	___	___	___	*sxwa*
KAXET	*črdili*	*zurgi*	*sc̣ore*	*ǯvari*	*ʒma*	*sxwa*
KIZIQ	*črdili*	*zurgi*	*sc̣ore*	*ǯvari*	*ʒma*	*sxwa*
FEREY	____	*zurgi*	*c̣ori*	___	*ʒma*	*sxo*
INGIL	___	___	*c̣oor*	*žor*	*zmaj*	___
MESH	*čyrdili*	*zurgi*	___	___	*ʒma*	*sxua*
RACH	___	*zurgi*	*c̣ori*	*ǯvari*	*ʒma*	*sxva*
H-IMR	*št̩ili*	*zrugi*	*st̩ori*	*ǯvari*	*ʒmai*	*sxva*
L-IMR	*št̩ili*	*zrugi*	*sori*	*ǯvari*	*ʒmai*	*sxwai*
LCHX	*štili*	___	*xc̣ori*	*ǯvari*	*ʒma*	___
ADJA	*čidili*	*ʒrugi*	*st̩ori*	*ǯvari*	*ʒma*	*sxwa*
TAO	*čirdili*	___	___	___	*zma*	*sxwa*
IMRX	___	*zrugi*	*sc̣ori*	*ǯwari*	*zmai*	*sxwa*
H-GUR	*čtili*	*ʒrugi*		*ǯwari*	*ʒmai*	*sxwai*
L-GUR	*čtili*	*zrugi*	*st̩ori*	*ǯwari*	*ʒmai*	*sxwai*

The output is striking, as compared to the previous set of dendrograms, obtained from a comparatively large amount of data. The dendrogram in Figure 2a is considerably more similar to the standard taxonomy (i.e. SCGD) for the Georgian Dialect Network than the previous one: two well balanced cores are now competing: a western complex (with GUR & IMR/LECHX as a kernel block, and ADJA, IMRX & TAO as a satellite, according to Bartolian CPE) versus a Central-North-Eastern core, in which GRG and KAR are properly embedded, while TUSH shows up, again, in the periphery as an outlier. The outlier cluster federates mostly eastern varieties, with a

northwestern core (XEV, MTIU/GUD, MOX) as opposed to two peripheral satellites: MESH in the westernmost southeastern stripe of the territory on the one hand, and INGIL on the other hand. This clustering points more at a default condensation of statistically asymmetric objects than to any close kinship. One should not forget that genealogy is but a by-product of automated language classification rather than a proper phylogenetic tool, as e.g. the cladistic method applied to biological entities may be. Above all, what quantitative dialectology produces is *statistical taxonomies* – which can eventually be interpreted from a genealogical standpoint.

Figure 2. Hissing vs. Hushing obstruents: a sample from the LaDyCa database

Figure 2a. Ward's method, Eight cognates	Figure 1b. Statistical distribution of differences

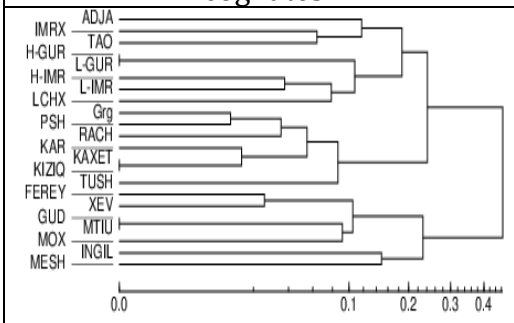
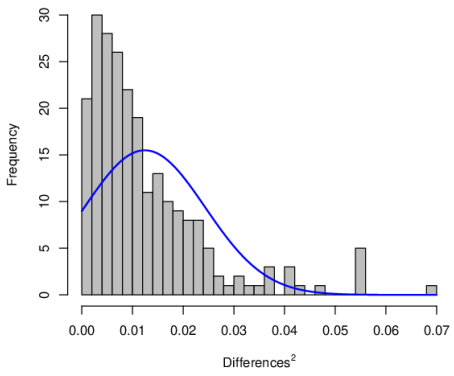
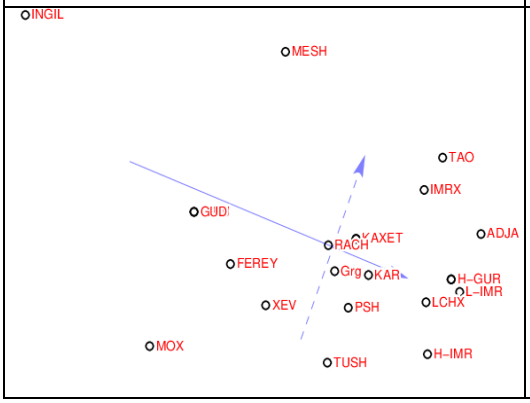
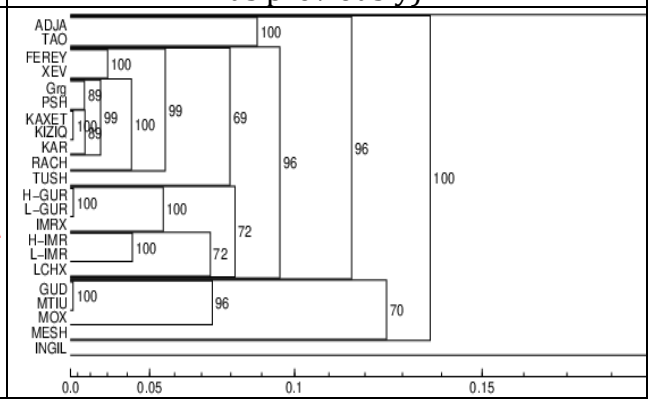

	

Figure 2c. Multidimensional scaling (r = 0,92)	**Figure 2d. Probabilistic dendrogram (Weighted average + Group Average; limit & as previously)**
	

However, the topologies of the various diagrams in Figure 2a-d differ substantially from the previous ones in Figures 1a-d. The statistical distribution (Figure 2b) is fairly similar, with negative skew and a similar threshold of complexity, so that this small sample of hissing & hushing obstruents can be considered as a fractal of the whole database. Instead, the Multidimensional Scaling Plot shows up as far more diffuse than in the previous figure, and most of the objects (i.e. varieties) tend to be more or less equidistant. Sections, separated by the two arrows (statistical vectors), are clear-cut, and the major central-eastern dialects, such as KAR, KAXET and of course Georgian (Grg) represent the centre of gravity. In Figure 2d, the same impression of consistency and distinct blocks is confirmed, and probabilistic estimations range very high, except for the most peripheral varieties (index 70). A low index of 69 also points at some

fuzziness at the root of the central block however -confirming the FPE for this complex, innovative and expanding area, at the very core of the GDN.

Therefore, the hissing-hushing obstruents variable might indeed deserve the status of variables of the DSL-type at GDN level, as they have already proven to be decisive at a much higher stage of phylogenetic taxonomy, for Kartvelian languages. Nevertheless, the effect is somewhat smoothed over by other factors, like VB (Variable Bleeding), in labialized contexts such as for items 2, 4 (in Table 1), 12 and 14 (in Table 2), and a strong Bartolian CPE, which has strengthened the central-eastern variants throughout the GDN over time. Nevertheless, some noise entered in the processing of the variable, as other variables (such as the endemic *-i* stem extensions, vowel laxity, etc.) were not excluded, in order to cling to "real", raw data. In the next set of results, I will make sure this noise does not interfere, in order to test its effect on the output.

3.2. The uvular stops variable

Uvular stops, simplex and ejective (as in GRG *çq̇ali* 'water', XEV ***q****orci* 'meat' vs. GRG ***x****orci*) also range among the dialectically relevant variables to analyze. The next set of results mostly concerns uvular onsets, in strong position, free from all other variables.

Figure 3. Uvular stops, simplex and ejective.[12] Results from *Gabmap*.

Figure 3a. Ward's Method, 5 clusters

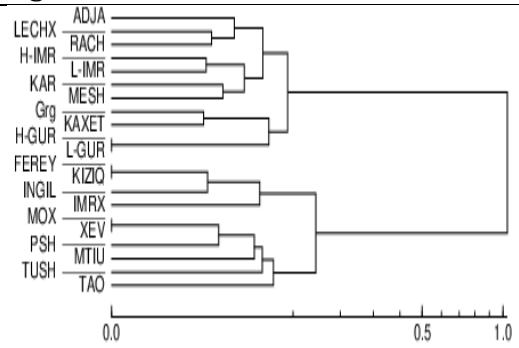


Figure 3b. Difference map

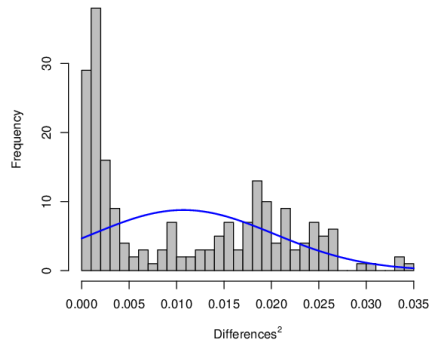


Figure 3c. MDS plot (r = 0,98)

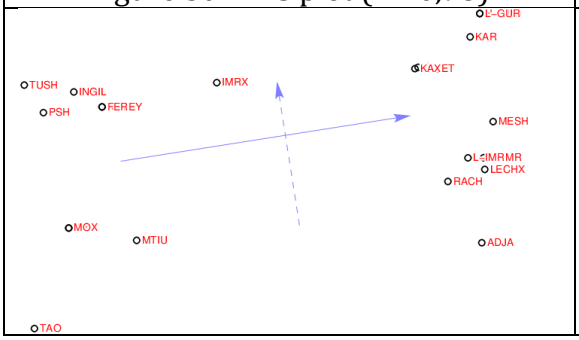
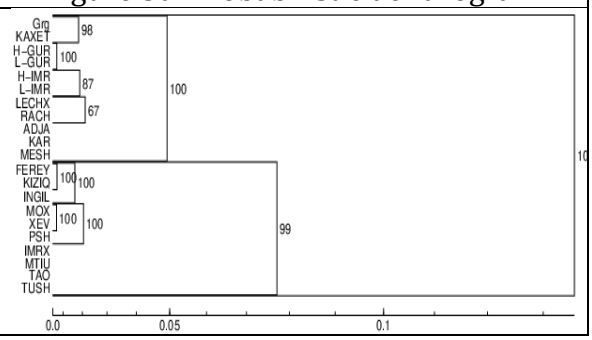


Figure 3d. Probabilistic dendrogram



This topology is nevertheless rather fuzzy, although patterns are clear-cut (see Figure 3c: a much scattered MDS plot, with no centre of gravity, and bundles of more or less equidistant objects). The East-Centre vs. West division is blurred (KAR is

[12] Size of the sample: 21 varieties, 10 cognates/items, 210 instances, 1068 characters (23 unique), 1025 tokens (24 unique).

included in the Western cluster; Grg and KAXET mingle with GUR, in the upper clade of Figure 3a). The core of the diasystem is missing here, as in the MDS plot. The lower clade in Figure 3c does indeed capture the "easternness" of the diaspora varieties of FEREY (in Iran) and INGIL (in Azerbaijan), for the first time in our dendrographic survey, associating them with KIZIQ, but the fact that these varieties can cluster here with a south-western subdialect as IMRX is embarrassing. Moreover: the fact that KIZIQ is separated from KAXET, in two different major clades, is simply a serious flaw. The lower clade is a trifle (but not much) better: the clustering of all north-eastern varieties in the kernel of the clade is good news -with TUSH as a satellite in this microtopology, as could be expected. But the fact that this kernel clusters with such a far distant south-western variety as TAO sounds like bad news, even if it shows up here as a default eccentric object. In short, not much can be salvaged from this experiment of trying to get the purest of an allegedly heuristic variable -the uvular stops-, and of suppressing the "noise" induced by the occurrence of other variables in the subset of data. In fact, anyone familiar with uvular stops (and uvular ejectives) in the world's languages would have suspected this result. The explosive phase in these phonological segments is so prone to allophonic variation that one should not bet too much on them for diasystemic taxonomy.[13]

Therefore, the characterization of the uvular stop & ejective variable can by no means pretend to qualify as a DSL variable (a major split variable). It doesn't fit the FPE (the Feature Pool), nor the Bartolian CPE; it is partially conditioned by VB (contextual variation, relative chronology), and the nice clustering of the north-eastern varieties in the lower clade of Figure 3a suggests that, at best, it can be involved in the EI layer of emerging isolates (or subdialects).

3.3. The labialized stops variable

I will further examine another phonological variable which I would place at the crossroads between the efficient and heuristic hissing-hushing variable and the uvular plosive & ejective: labialized stops (the Cv- variable),[14] as in Table 4.

Table 4. A sample of the Cv- variable, from the LaDyCa database

	15	**16**	**17**	**18**
	'still, yet'	**'child, son of'**	**'you' 2Pl**	**'eye'**
GRG	*ḳidev*	*švili*	*tkven*	*tvali*
XEV	*ḳide*	*švili*	*tkven*	*tvali*
PSH	*ḳiden*	*švili*	*tkven*	*tvali*
MOX	*ḳidav*	*švili*	*tkwen*	*toli*
MTIU	*ḳidan*	*švili*	*tkven*	*twäli*
GUD	*ḳidav(a)*	___	*tkven*	*toli*
TUSH	*ḳiden*	___	___	*tvalĭ*

[13] See http://aldelim.org/ and in particular http://dalima.aldelim.org/ for Eastern Mayan languages. We also noted much allophonic variation between the plosive and the fricative realization of uvular stops and ejectives in Totonaco-Tepehua doing fieldwork in 2015 in the state of Puebla, Mexico.

[14] Data processed: 22 varieties, 10 cognates, 220 instances, 1124 characters (29 unique), 1104 tokens (29 unique).

KAR	*kide*	*švili*	___	___
KAXET	*ḳidena*	*šwili*	*tkwen*	*twali*
KIZIQ	*ḳidena*	*šwili*	*tkwen*	*twali*
FEREY	*ḳide*	*švili*	*ken*	*toli*
INGIL	*ḳedem*	*šül*	*tkön*	*tol*
MESH	*ḳide*	*šüli*	*kten*	*tvali*
RACH	*ḳidomec*	*švili*	*tkven*	*tvali*
H-IMR	*ḳide*	*švili*	*tkven*	*twali*
L-IMR	*ḳido*	*švili*	*tkven*	*twali*
LCHX	*ḳidom*	*švili*	*tkvene*	*twali*
ADJA	*ḳido*	___	___	*twali*
TAO	*ḳide*	*šwili*	*tkwen*	___
IMRX	*ḳido*	*šwili*	*tkwen*	*twali*
H-GUR	*ḳido*	___	*tkven*	*tvali*
L-GUR	*ḳidom*	___	*tkven*	*twali*

This phenomenology is interesting in several respects. Although it does not qualify as congruent with the standard taxonomy of Georgian dialects on several decisive points (in the upper clade, western ADJA and GUR mingle aberrantly with eastern varieties such as KAXET and KIZIQ), many details of the inner structure of the topology in Figure 4a provide interesting cues about the general, or perhaps even the deep structure of the GDN.

Figure 4. Labialized stops in the LaDyCa database: a sample. Results from Gabmap.

Figure 4a. Ward's Method

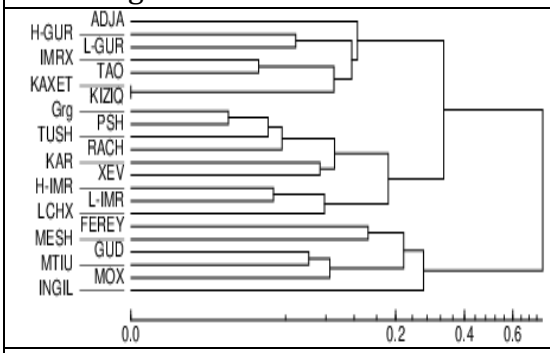


Figure 4b. Difference map

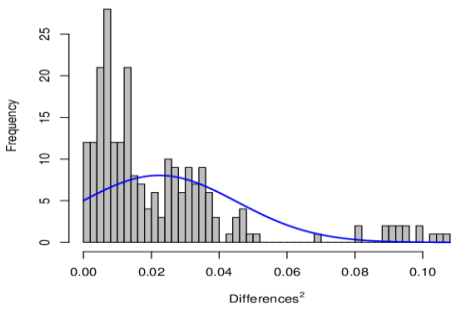


Figure 4c. MDS plot

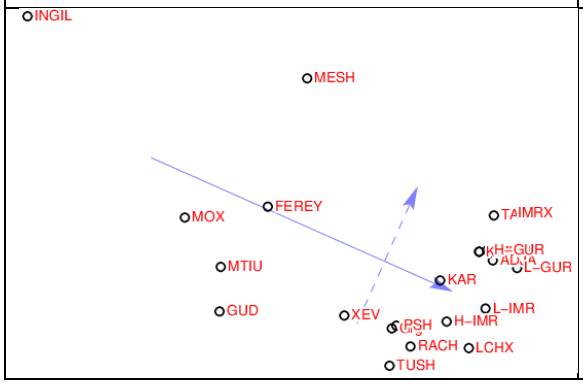
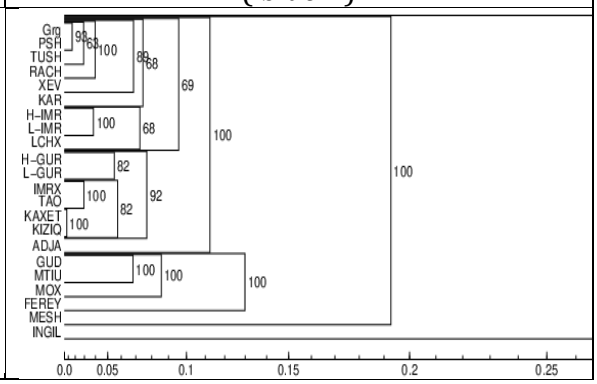


Figure 4d. Probabilistic dendrogram (*ibidem*)



First, there is a solid core in this topology: the central clade, which associates the central (GRG, KAR) and the north-eastern varieties (TUSH, PSH, XEV, with the addition of RACH in the West, in accordance with a northern highland continuity dynamic)[15] as a core, on the one hand, and the IMR/LCHX dialect complex in the North-West on the other hand. This is very consistent, and it points clearly at the kernel, or the nucleus of the DSL (the major division between C/E and W), and at the Bartolian dynamics –both areal cores having been or being leaders in their geographic realm. Second, the clade in the lower part of the dendrogram indicates two patterns: on the one hand, a default clustering of peripheral eastern dialects (FEREY, MESH, INGIL), on the other hand, a split of the north-eastern bundle of varieties, since MTIU/GUD and MOX join them in this clade –i.e., the westernmost complex out of this compact bundle. Now, remember we saw the same split in the previous topology (uvular stops, previous section). One could think this was due to erratic behaviour of the sample. We see it may not be so. These facts, including the mingling of RACH with the easternmost varieties of the north-eastern complex in the northern highlands, might hint to some kind of a former "cordillera" dialectical continuity, which somehow attempted to emerge at some point in history, but collapsed thereafter. Third, if the chain MTIU/GUD-MOX in the centre of the lower clade can be seen as a core area, then the FEREY & MESH subclade qualifies as a first concentric circle of default aggregation, whereas INGIL comes next, in a second circle of distant, although plausible convergence. INGIL is no longer a satellite disconnected from the rest of the GDN: its structural ties with the eastern complex surface clearly for the first time. Although this set of results might seem arguable because of some mingling between East and West, it does bring new and interesting cues about the inner structure of the GDN.

Moreover, the distribution of differences in Figure 4b differs from the previous one: it is less negatively skewed and looks more like a Gaussian curve than in the previous difference maps. The MDS plot points at a FPE trend, with most varieties condensating in a pole, at the right of the plot, while KAR stands not far from the centre of gravity, as could be expected. What strikingly emerges from this picture is a kind of central-western configuration, instead of the canonical central-eastern confederation. Escaping from this cloud, the "northern cordillera" and the easternmost peripheral varieties rise to less crowded areas of the plot: to the left for MOX and MTIU/GUD, and far above, as far as MESH and INGIL are concerned.

The main conclusions we can draw from section 3 are as follows, in (2)

(2) Trends in the GDN according to the GMDD

(2a) Although the GDN started with a set of structural variables inherited from the various phases of the Kartvelian split, it has long undergone a process of bipolarization between the core of the Western block (IMR/LECHX, and GUR as a first circle periphery) and the core of the Central-Eastern block (mostly KAR). Each of these cores developed fairly strong Bartolian CPE.

(2b) The BZE (Buffer Zone Effect) in the centre (involving the north-eastern complex, in strong interaction with KAR) involves a strong FPE I (Feature Pool) with eastern Georgian (KAXET/KIZIQ).

(2c) At the periphery of these competing cores, two major trends arise: either EI (emerging isolates) from the second or third Bartolian circle (CPE), such as PSH,

[15] I would even use the term, in this case, of a kind of "cordillera effect", as the variable here follows the Transcaucasian mountain ridge.

TUSH in the north-eastern zone, or ADJA, IMRX in the West, out of VB (variable bleeding) and PLEP (Phonolexical Endemic Patterns) on the one hand, or straightforward singleton dialects out of language contact beyond the frontiers of the two former geopolitical hegemonies (Turkey and Persia), on the other hand.

(2d) Differences have been eroded by the powerful influence of the literary language, although the complex shape of the geolinguistic landscape has to some extent preserved retentive trends, as in the Greater Caucasian highlands, from RACH to TUSH, or in the valleys where ADJA is spoken. Some varieties such as MESH or TAO have emerged from intricate processes of settlement and resettlement, and are embedded in contact with exogenous languages, such as Armenian for the former, and Turkish for the latter. All these peripheral zones show greater idiosyncrasy, oscillating between the CPE and the EI. Consequently, they are preserved from the FPE (Feature Pole Effect) or from the BZE (Buffer Zone Effect).

(2e) Due to the phonological and morphological complexity of Kartvelian languages, VB (Variable Bleeding, or Relative Chronology) and PLEP (Phonolexical Endemic Patterns) should be carefully sequenced and controlled in any quantitative attempt to measure distance or similarity in the GDN. These two effects contribute to the blurring of deep evolutive trends in the GDN (as in Figure 1, with 243 cognates). In processing phonological variables, we tried to avoid this bias. This selective approach gave encouraging results, though partial and, of course, provisional.

(2f) In the future, we can predict that the GDN will intensify a monopolar Bartolian Dynamic Field (CPE), with GRG in the lead. Yet, competing core or first-circle peripheries such as GUR may still flourish to some extent. Most second or third circle varieties will either fade away under this assimilation process by Standard Georgian, or will survive in their foreign community surroundings, although some, such as TAO, are now highly endangered (Chokharadze *& al.* op. cit.).

Therefore, an urgent task is to observe these processes of dialect assimilation or resilience and transformation, and to document as many oral and written texts as possible (as *The Corpus of Georgian Dialects*-project already does, see Beridze *& al.* 2009). Moreover, particular attention should be paid to oral history of interactions between dialects, within the framework of the South Caucasian Vertical archipelago (Nichols, *op. cit.*) and its fractal components (e.g. the North-Eastern complex in relation to KAR; ADJA inner complexity, in relation to TAO and to IMRX, as in Chokharadze *& al.* ibidem).

4. Conclusion and prospects

What did this survey of a 243 cognates of Georgian teach us that we did not already know about a well known topic? First, that standard taxonomies should not be taken as definite knowledge and unquestionable authority: most of their power lays instead in their refutability and falsifiability. Through the process of challenging the canonical picture of a dialect network with statistical tools, we can grasp otherwise unattainable cues on the complex inner structure of the diasystem. Second, by challenging the qualitative picture of the dialect network, quantitative methods also face a challenge. The filiation of emerging isolates (EI) such as INGIL or FREREY is not easily captured by quantitative methods, whereas it is more easily established by philological and qualitative

tools (isoglosses). Yet, it depends on how we calibrate data samples. In turn, the quantitative approach may enhance deep patterns which could hardly be unearthed with qualitative tools, such as the "northern cordillera effect". Second, quantitative tools are designed for calculation and taxonomy building - a noble yet trivial task, epistemologically speaking. We need more, for the sake of General Dialectology. Quantitative tools, such as computational dialectology, are all the more powerful if they are rooted in linguistic theories, such as the Gudschinsky's Model (the GMDD). From this standpoint, the GDN continues ancient trends which have been decisive in the Kartvelian linguistic stock, such as the hissing-hushing correlation, which point at the DSL. From the Western versus Eastern split, a complex buffer zone (the BZE, or Buffer Zone Effect) has emerged, bringing together the Kartlian-Meshkhian area (KAR & MESH), in the midlands, with the North-eastern area, in the north-western highlands (MOX, MTIU/GUD, XEV, PSH, TUSH). This interaction, of the Vertical Archipelago type (Nichols 2008), has weakened a plausible former Transcaucasian highland continuum, from TUSH to RACH. Eastern and central dialects have mingled conspicuously over time, so that the eastern dialect complex has constantly been expanding (with transborder outliers such as INGIL and FEREY) and recessing (being incorporated by KAR, or developing into MESH, migrating towards the South-West). This constant diastole and systole movement often happened under pressure from foreign hegemonies, especially the Persian hegemony, to the East (EF), and gave shape to a dense Feature Pool (FPE). However, the GDN has also evolved according to a Bartolian dynamic field (CPM), opposing two main innovative dialects: KAR vs. IMR. In turn, Western peripheral dialects, especially GUR, have more or less followed the trends expanding from the North-West (from IMR), and could even go beyond (causing VB entropy). From the South-West, the trend has rather been peripheral idiosyncratic innovation (EI), and diastole, heading westward, leading to the emergence of innovative varieties such as ADJA, IMRX and TAO.

The overall shape of the GDN is quite unified, as the different maps suggest, and the low differences to be found in the matrix of data in Appendix 1.1-2. Standard Georgian, as a powerful and prestigious dialect, with a long-time written tradition, and outstanding tools for diffusion (among which a very original and practical alphabet and spelling), has been unifying the dialect network for a very long period already. Yet, underneath, a rich array of fine-grained variation still lingers and provides many hints at history and at original patterns in the phonology, the grammar and the lexicon of this most valuable language - Georgian.

References

ASATIANI Rusudan, 2008, Directions of Diachronic Developments of the Kartvelian Languages, paper presented at the *44th International Congress on Medieval Studies, University of Kentucky*, available Online: https://www.ayk.gov.tr/wp-content/uploads/2015/01/ASATIANI-Rusudan-DIRECTIONS-OF-DIACHRONIC-DEVELOPMENTS-OF-THE-KARTVELIAN-LANGUAGES.pdf.

BARTOLI Matteo, 1945, *Saggi di linguistica spaziale*, Turin, Rosenberg & Sellier.

BERIDZE, M. *& al.*, 2009, The Corpus of Georgian Dialects. In *Proceedings of the NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference Smolenice*, p. 25-35, *Jazykovedný ústav Ľ. Štúra SAV*, Bratislava, Slovakia.

GAMKRELIDZE Th., 2005, Reconstruction of proto-language and its chronological levels, *Perspektiva*-XXI, VI, Tbilisi, ena da kultura, p. 84-87.

GIGINEISHVILI Ivane, TOPURIA Varlam & KAVTARADZE Ivane 1961 *kartuli dialekt'ologia I* [Georgian dialectology, I], Tbilisi, TSUG.

GUDSCHINSKY Sarah C., 1958, Mazatec dialect history: A Study in Miniature, *Language* 34, p. 469-481.

CHOKHARADZE Malkhaz, RODINADZE Shota, ABASHIDZE Natia, 2018, The Taoan Dialect of Georgian Language and Linguistic Space of Collective Memory in the Chorokhi Basin, *Journal of Arts and Humanities*, 07-10, p. 65-72.

HEINSALU Els, PATRIARCA Marco, LÉONARD Jean Léo, 2020, *Languages in Space and Time. Models and Methods from Complex Systems Theory*, Cambridge University Press.

KASSIAN Alexei, 2015, Towards a formal genealogical classification of the Lezgian languages (North Caucasus): testing various phylogenetic methods on lexical data. *PLoS ONE* 10(2): e0116950. doi:10.1371/journal.pone.0116950.

LEONARD Jean Léo, 2019, Méthodes pour l'analyse et la documentation des langues et des dialectes kartvèles, à la lumière de la Théorie des Dynamiques Langagières / Théorie de la Complexité. In MIKADZE, Manana *& al.* 2019. *The Fifth International Scientific Conference Language & Culture*. Kutaisi, Akaki Tsereteli State University, p. 625-650.

LEONARD Jean Léo, 2017, Le projet LaDyCa (*Language Dynamics in the Caucasus*, IDEX Emergence 2017-18) : avancées et résultats, *Kadmos* 9, Tbilisi, ILIAUNI, p. 42-77.

LEINONEN Therese; ÇÖLTEKIN, Çağrı; NERBONNE, John 2016 Using Gabmap, *Lingua* 178, p. 71-83.

LEVENSHTEIN Vladimir I., 1966, Binary Codes Capable of Correcting Deletions, Insertions, and Reversals, *Soviet Physics‐Doklady*, 10/8, p. 707-710.

MACHAVARIANI G. 1965, *Proto-kartvelian consonants system*,Tbilisi, TSU press.

MUFWENE Salikoko, 2013, *Complexity perspectives on language, communication, and society*, in Massip-Bonet, Àngels & Albert Bastardas-Boada (eds.), 2013. *Complexity Perspectives on Language, Communication and Society*, Springer, p. 197-218.

MUFWENE Salikoko 2001, *The Ecology of Language Evolution*, Cambridge, Cambridge University Press.

MURRA John Victor, 1956, *The Economic Organization of the Inca State*, Ph.D. diss., Department of Anthropology, University of Chicago.

MURRA John Victor, 1985, El Archipielago Vertical Revisited, in *Andean Ecology and Civilization*, edited by Shozo Masuda, Izumi Shimada & Craig Morris, 3-13, Tokyo, University of Tokyo Press.

NICHOLS Johanna, 2004. The Origin of the Chechen and Ingush: A Study in Alpine Linguistic and Ethnic Geography, *Anthropological Linguistics,* 46-2, p. 129-155.

PICARD Flore, GÉRARDIN Hélène & LÉONARD Jean Léo, 2018, A report on the LaDyCaproject (*Language Dynamics in the Caucasus*, IDEX Emergence, 2917-18), in *Methods in theoretical and empirical dialectology, talk at Methods for Endangered Kartvelian varieties documentation & Analysis according to Language Dynamics Theory, II*, 23th of February 2018, Ilia State University, Tbilisi.

POPPER Karl, 1934, *Logic der Forschung*, Wien, Springer.

POPPER Karl, 1963, *Conjectures and refutations,* London, Routledge.

TUITE Kevin 1989. The geography of Georgian *q'e. The Non-Slavic Languages of the USSR: Linguistic Studies,* H. Aronson (ed.), Chicago Linguistics Society, 283-302.