LINGUISTICA URALICA LVII 2021 2 https://dx.doi.org/10.3176/lu.2021.2.01

IRINA NOVAK (Petrozavodsk),* *MARTTI PENTTONEN* (Kuopio)

ANALYSIS OF KARELIAN DIALECT DIVISION BASED ON ALGORITHMIC CLUSTERING

Abstract. The article presents an algorithm for clustering dialects by similarity of data in the dialect atlas of the Karelian language (Бубрих, Беляков, Пунжина 1997). By repeating the procedure we get a hierarchy of dialects. Cluster hierarchies can be based on all maps of the Atlas, or on any subset of maps, e.g. morphology, noun inflection, or vocabulary maps. As an example, we consider clusters based on sibilants, local cases, and all maps of the Atlas. An analysis of the clusters, with reference to linguistic literature, leads to the following conclusions: Karelian dialects can be divided into two main areas, Karelian Proper and Livvi-Ludic areas. Border Karelian dialects, which are relatively similar to each other, seem more like Livvi Karelian than Karelian Proper. On the other hand, the traditional volost based division of Karelian dialects turns out to be too fine-grained to reveal any significant differences.

Keywords: Karelian language, dialect division, algorithmic clustering.

1. Introduction

According to the overviews by Leskinen (1998 : 355—359) and Sammallahti (1977), the emergence of the Old Karelian language should be dated to the latter half of the first millenium of the current era, and located on the south-western coast of Lake Ladoga, Korela Karelia. Important influences behind the development of the Karelian culture and language have been the Viking era, the rise of the Swedish and Russian states and their wars in Karelia, the era of Swedish dominance, and the era of Russian dominance. In the background of the current language situation are the migrations of Karelians in the 17th and 18th centuries and after the World Wars I and II.

The formation of the main dialect called Karelian Proper was due to the migrations of Karelians from the western coast of Ladoga and current Finnish North Karelia to the north and to the Tver area in central Russia, as a consequence of the 17th-century wars (Leskinen 1998 : 358—359; ALFE 1 : 8). The migrations of the Karelians to the east and their contacts with Vepsians on the Olonets Isthmus led to the development of two main dialects of the Karelian

* The study by Irina Novak was carried out under state order of Karelian Research Centre of the Russian Academy of Sciences.

language: Livvi (Olonets Karelian), in which the Karelian component prevailed, and Ludic, which has a strong Vepsian component (Itkonen 1971 : 179, 182).

By definition (JIEC 1990), main dialects are divided into dialects and further into subdialects. Dialect is a language form used for communication by people in a close territorial community (in the case of the traditional classification of the Karelian language, we are talking about volosts). Dialects differ in sound system, grammar, word formation, and vocabulary. Subdialect is the smallest territorial variety of language used by residents of one or more neighboring settlements not having essential territorial linguistic differences.

Currently there is no agreement in the linguistic research community upon Karelian language division. For example, in the Russian research tradition, Ludic is considered to be one of the main dialects of the Karelian language (Зайков 1999 : 7) (see Figure 1). In the Finnish research tradition, however, Ludic is considered to be either a mixture of Karelian and Vepsian dialects (Leskinen 1998 : 382), a dialect of the Vepsian language (Genetz 1872), or a distinct language of its own (Pahomov 2017). According to Jeskanen (2019 : 3—4), Karelian consists of four language varieties: Viena, Karelian (Karelian Proper, southern dialects, in terms generally accepted), Livvi, and Ludic. Jeskanen calls them all languages.

Since the 1870s, Karelian dialects have been studied by Finnish (Genetz 1872; 1880; 1885; Ojansuu 1907; Turunen 1946; 1950; Itkonen 1971; Leskinen 1998; Virtaranta 1972) and Russian (Бубрих 1947a; Беляков 1958; Зайков 1987; Рягоев 1993) linguists, but many unsolved problems still remain as for dialect division. The problem of terminology: Are dialects or subdialects correctly represented on the traditional map of Karelian dialects? The problem of choosing the basic principle of dialect division: Are administrative borders adequate for representing the dialectal division of the Karelian language? The problem of determining the status of individual groups of dialects: What is the position of, e.g., Ludic, Border Karelian, or transitional dialects?

In this work we approach the dialect division in a new way by using a graph algorithm in order to group the dialects into hierarchical clusters, using the data in the dialect atlas of the Karelian language (Бубрих, Беляков, Пунжина 1997; hereinafter: Atlas). The clustering method presented in this article was developed in parallel with the project "Karelian Language in Grammars" (2015—2019, supervised by Lea Siilin) and widely applied to the Atlas maps. Results can be found either in the monograph Novak, Penttonen, Ruuskanen, Siilin 2019 or on the project website http://karjalankieliopit.krc.karelia.ru. Note, however, that the presentation of the algorithm was beyond the scope of the book. Recently a different clustering algorithm has been applied to related problems (Lehtinen, Honkola, Korhonen, Syrjänen, Wahlberg, Vesakoski 2014; Honkola, Santaharju, Syrjänen, Pajusalu 2019). We believe that for our purpose and the Atlas data, the parallel spanning tree-based clustering algorithm is more suitable and intuitive.

2. Problems of dialect division

One of the most difficult questions in Finnic linguistics is the classification of Ludic dialects, which is mainly due to the differences in Russian and Finnish research traditions. According to D. V. Bubrich (Бубрих 1947a; 1948), Ludic dialects belong to the Karelian language (Зайков 1999; Atlas). In another view, generally supported by Finnish researchers, Ludic is a distinct language (KKVS;

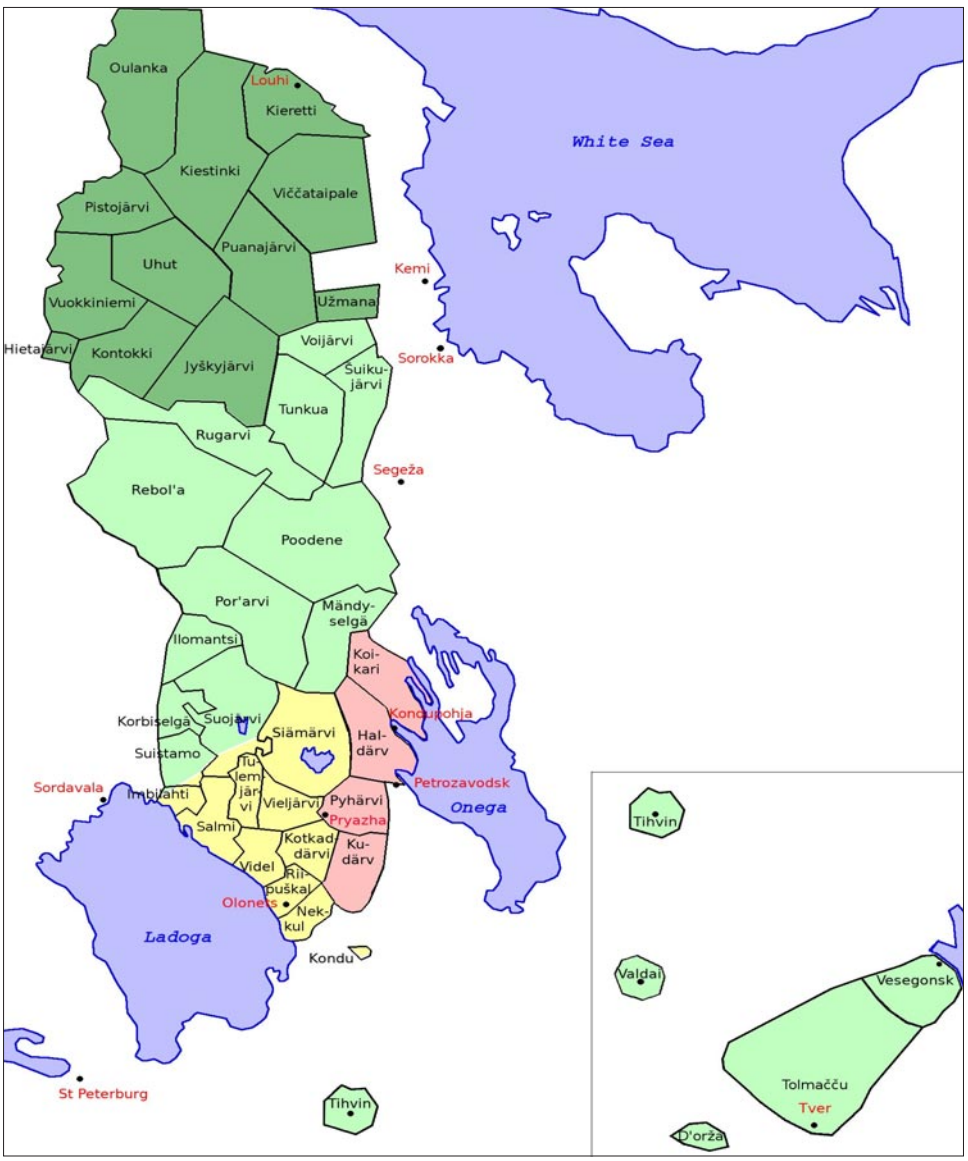

Figure 1. **Map of the traditional dialect classification of the Karelian language.**

Virtaranta 1972; also Pahomov 2017 : 9). There has also been dialogue about the position of Border Karelian dialects: Do they belong to Karelian Proper or Livvi Karelian, or are they mixed dialects (Uusitupa, Koivisto, Palander 2017; Koivisto 2018 : 59, 71)?

A number of questions have been raised concerning the classification of dialects by the volost division of the Karelian territory in the early 20th century. In the preface of the Atlas (Керт, Рягоев 1997 : 4, 8) it is stated that "within the volost at that time, economical and cultural life were closed, which could not but affect the language". This classification was developed in the 1980s and1990s while preparing the relevant linguistic atlases (ALFE 1; Atlas). It was chosen as the basis for dialectological research in Зайков 1999 and Зайков 2000, and has since become widely accepted.

The map in Figure 1 presents a combined view of the Finnish (KKS 6) and Russian (Зайков 1999 : 7) research traditions of dialect classification. They differ in the scope of the dialects. The Finnish dialect map does not cover Ludic dialects as they are not considered to be dialects of the Karelian language. The Russian map does not cover Border Karelia as the Karelian population was evacuated to Finland due to the 1939—1944 wars between Finland and the Soviet Union.

It should be noted that it is not always possible to detect significant differences between neighbouring dialects in the traditional classification based on the volost division of the territory of Karelia. Such is the case with some Viena Karelian dialects of Karelian Proper, and the Nekkul and Riipuškal subdialects of Livvi Karelian. On the other hand, significant differences exist in phonetics and morphology between some subdialects, even though they are classified as belonging to the same dialect (see the diphthong system and the formation of plural forms in the Poodene dialect, for example.

It should also be noted that isoglosses of dialect features do not always coincide with volost boundaries. For example, P. Virtaranta mentions the dialect of Miinoa village in his article "Die Dialekte des Karelischen" (1972). The village was located in Rugarvi volost (in the west, close to Kontokki), but linguistically it is closer to the Kontokki dialect than to the Rugarvi dialect. So is the case with the Southern Suojärvi subdialects that are closer to Livvi Karelian than to the South Karelian dialects of Karelian Proper (Virtaranta 1972 : 10). An analysis of Ludic subdialects showed that the dialect of Kaškana village is significantly different from the Kuďärv dialect, although these villages are administratively related, while the Kaškana dialect is not so different from the neighboring South Ludic subdialects (Wiik 2004 : 30—31).

3. Dialect atlas of the Karelian language

The Atlas (Бубрих, Беляков, Пунжина 1997) contains a vast amount of dialect material. This is a pioneering work and an invaluable resource for researchers of Karelian dialects. The collection of the data was guided by a question booklet containing about 2,000 dialectological questions (Бубрих 1937; updated 1946). The first hundred maps of 150 Karelian localities were compiled already in 1937. By 1948, about 800 test maps had been drawn up. After the death of Bubrich, the work on the Atlas was continued by A. A. Beljakov, with the participation of N. A. Anisimov and G. N. Makarov, together with editors N. I. Bogdanov and M. M. Hämäläinen. The work was completed in 1956. However, for a number of reasons the Atlas remained unpublished. In 1956—1958, under the leadership of G. N. Makarov, question booklets were filled in for 60 settlements of the Kalinin region. In 1970—1973 A. V. Punžina filled in the question booklets for six more points, and in 1990—1991 she prepared maps for Tver Karelian dialects. However, not all settlements were included in the Atlas when it was published in Finland in 1997, edited by L. Sarvas (Керт 2002 : 24—38). The final version contained 209 maps divided into sections on "Language information" (maps 1—3), "Phonetics" (maps 4—112), "Morphology" (113—185), and "Vocabulary" (186—209). Out of all the material collected during the work,[1] 186 settlements (150 of Karelia) and (36 of the Tver region) were

[1] Materials are stored in the Scientific Archives of Karelian Research Centre of Russian Academy of Sciences (fund 1, inventory 38).

selected for the final version of the Atlas. It should be noted that only those maps were selected for publication that show the greatest number of differences in phonetics, grammatical categories or lexemes, on the level of dialects or main dialects of the Karelian language.

The first effort to use the whole of the Atlas for dialect classification was made by the Finnish linguist K. Wiik. In 2004, he published the report "Quantitative research of Karelian dialects" (Wiik 2004), in which he used the Atlas in a novel way. For each of the 186*185/2 pairs of settlements, he counted in how many maps they have different variants. By using the difference table he reasoned, by percentage numbers, how much the dialects of the Karelian language differ from each other. Thus it was possible to establish, for example, that the difference between Karelian Proper and the Ludic dialects is on average 76%, while the difference between Karelian Proper and Livvi dialects is 65%, and 54% between the Ludic and Livvi main dialects (Wiik 2004 : 63—64, 67). Based on these tables, the boundaries between the groups of dialects and subdialects of the Karelian language were outlined. The boundaries proposed by Wiik differ significantly from those of the traditional classification: the traditional dialects of Viena Karelian (excluding Vičča̧taipale and Užmana) were combined into one dialect; so were Tunkua and Šuikujärvi, Ludic Pyhärvi and Kuďärv also formed single dialects; whereas in Tver Karelia three dialects were replaced by five.

4. Dialect data and algorithmic clustering of dialects

We continue where Wiik left off. He used a computer to store the Atlas data and compactly present the differences in table form for further analysis and conclusions. Instead of human selections and conclusions, we use a computer algorithm to objectively group dialects into clusters based on the smallest number of differing features in the Atlas.

At the first stage of the work, we rewrote the Atlas data in a suitable form. Each of the maps from 4 to 209 was rewritten as a file with some title data and a line for each settlement (or point of the map). There are 150 + 36 points in the Atlas, but we added six more points to represent Border Karelian. True, Border Karelian data is not Atlas material as it was collected from six dialect samples from Leskinen 1934. Due to the very narrow samples of the dialects, not all Atlas questions could be answered, and hence the results concerning this area are not as reliable as the results concerning other areas. No data at all was available about the Tihvin and Valdai dialects of Karelian Proper, but we hope to make relevant additions in the future. The dialect variant at each point was encoded by a letter a,b,c,..., i.e. comma-separated letters if many variants were found, or by '–' if no data was available. Hence, there were 192 × 206 records of dialect variants.

For the analysis of the Atlas data, we used an algorithm which in computing science literature is known as Sollin's minimum spanning tree algorithm. A form of this algorithm was first published already by O. Borůvka (1926) for the optimal design of electrical networks, and again by G. Sollin (1965) for similar purposes. In the original application, the goal was to design a network so that the wire length is minimal. For our purpose, a suitable form of the algorithm is a parallelized version which proceeds "bottom-up", level by level (Jájá 1992), as dialects also develop in parallel. First, the points are connected

to their nearest neighbours, forming groups that we call clusters. On higher levels of clusterization, clusters are connected to the nearest clusters to form larger clusters, and so on, recursively, until all nodes belong to one big cluster. Cluster hierarchy is analogous to the dialect hierarchy, rising from the level of subdialects to dialects, main dialects, and languages.

Verbally, a parallelized version of Sollin's algorithm, applied to dialects, can be written as follows:

Clusterization of Karelian dialects

1. Let the points on the geographical map be $p_1, p_2, p_3, \ldots$ and the numbers of the thematic maps be $m_1, m_2, m_3, \ldots$ Let the answer at point p_i to the question on map m_k be a_{ik}.
2. For all pairs of map points p_i and p_j, let d_{ij} be the number of maps m_k such that $a_{ik} \neq a_{jk}$.
3. From each point p_i draw an arrow to the point p_j for which d_{ij} is smallest. These arrows connect points to groups called clusters. One can prove that each cluster has exactly one pair of points whose arrows point to each other. One of these two, whichever occurs earlier in the list of points, is called the root of that cluster.
4. In each cluster, redirect the arrows of the points to the root of that cluster.
5. If there are more than one cluster, construct a new difference table for the roots of the clusters as follows: If p_i and p_j are roots of clusters, define the difference d_{ij} of clusters (with roots) p_i and p_j to be the smallest d_{xy} such that point p_x is in cluster p_i and point p_y is in cluster p_j, mathematically
$$d_{ij} = \min\{d_{xy} \mid p_x \text{ is in cluster } p_i \text{ and } p_j \text{ is in cluster } p_j\}.$$
6. Continue to step 3.

By counter assumption, it is easy to prove that whenever a connecting arrow is selected, there is no better way to connect those components (see Jájá 1992). Note that step 4 is purely cosmetic. All points are equal members of the cluster, the root is just the representative of the points in this cluster, and the purpose of the redirection of arrows is only to make the picture of the cluster more readable.

Example. Choose points 1, 4, 47, 57, 72, 80, 187, 188, 189, 190 and maps 4, 22, 63, 74, 78, 107, 121, 155, 161, 170, 186, 204 of the Atlas. The alternatives on map 4, for example, are a = *mua*, b = *muo*, c = *moo*, d = *moa*, e = *maa*, f = *mia*.

Data from the Atlas:

	4	**22**	**63**	**74**	**78**	**107**	**121**	**155**	**161**	**170**	**186**	**204**
1	b	c	a	a	b	a	a	c	a	a	d	a
4	a	e	a	a	b	a	a	c	d	b	df	a
47	a	c	a	a	b	b	a	c	b	a	d	a
57	a	e	a	a	b	b	a	c	d	b	f	b
72	d	b	b	a	a	a	b	c	d	d	–	a
80	a	b	b	a	a	b	b	a	e	d	a	c
187	a	c	a	a	b	a	a	c	b	a	d	a
188	a	c	a	a	b	a	a	–	b	–	a	d
189	d	b	a	a	b	a	a	a	b	c	ag	ad
190	d	b	ab	a	a	a	a	–	b	–	ad	ad

Difference table:

	1	**4**	**47**	**57**	**72**	**80**	**187**	**188**	**189**	**190**
1	–	4	3	7	7	11	**2**	4	6	4
4	4	–	4	**2**	6	10	3	4	6	4
47	3	4	–	5	8	9	**1**	3	6	4
57	7	**2**	5	–	8	9	6	5	8	7
72	7	6	8	8	–	5	7	7	6	**2**
80	11	10	9	9	**5**	–	10	7	8	5
187	2	3	**1**	6	7	10	–	2	5	3
188	4	4	3	5	7	7	**2**	–	2	3
189	6	6	6	8	6	8	5	2	–	**1**
190	4	4	4	7	2	5	3	3	**1**	–

For the nearest neighbours, see Figure 2a:

1	4	47	57	72	80	187	188	189	190
187	57	187	4	190	72	47	187	190	189

The first level clusters are (root first) as follows, see Figure 2b:

{47,1,187,188}
{4,57}
{189,72,80,190}

For second level clusters, we need to calculate the differences between the first level clusters. The difference between clusters 4 and 47, for example, is the smallest difference between a point in {4,57} and a point in {47,1,187,188}, i.e. pairs {(4,47), (4,1), (4,187), (4,188), (57,47), (57,1), (57,187), (57,188)}. Among these, (4,187) has the smallest difference 3. This has been drawn with a dotted line in Figure 2b. The whole difference table for of the first level clusters is:

	4	**47**	**189**
4	–	**3**	4
47	3	–	**2**
189	4	**2**	–

For the nearest neighbours, see Figure 2c:

4	47	189
47	189	47

For the second level cluster, see Figure 2d:

{47}

On the second level, we have only one root, which means that all points are in one cluster.

Note 1. We are liberal in calculating the differences. For example, a = '–' and a = a,b, but a ≠ b,c. In practice, a stricter rule a ≠ '–' etc. would not essentially change the clusters.

Note 2. Even though the root of a cluster may seem central, the main reason for choosing a root and draw lines from the other nodes of a cluster to the root is to make the algorithm deterministic. It is also visually motivated. No linguistic conclusions should be made about the "centre" of a cluster.

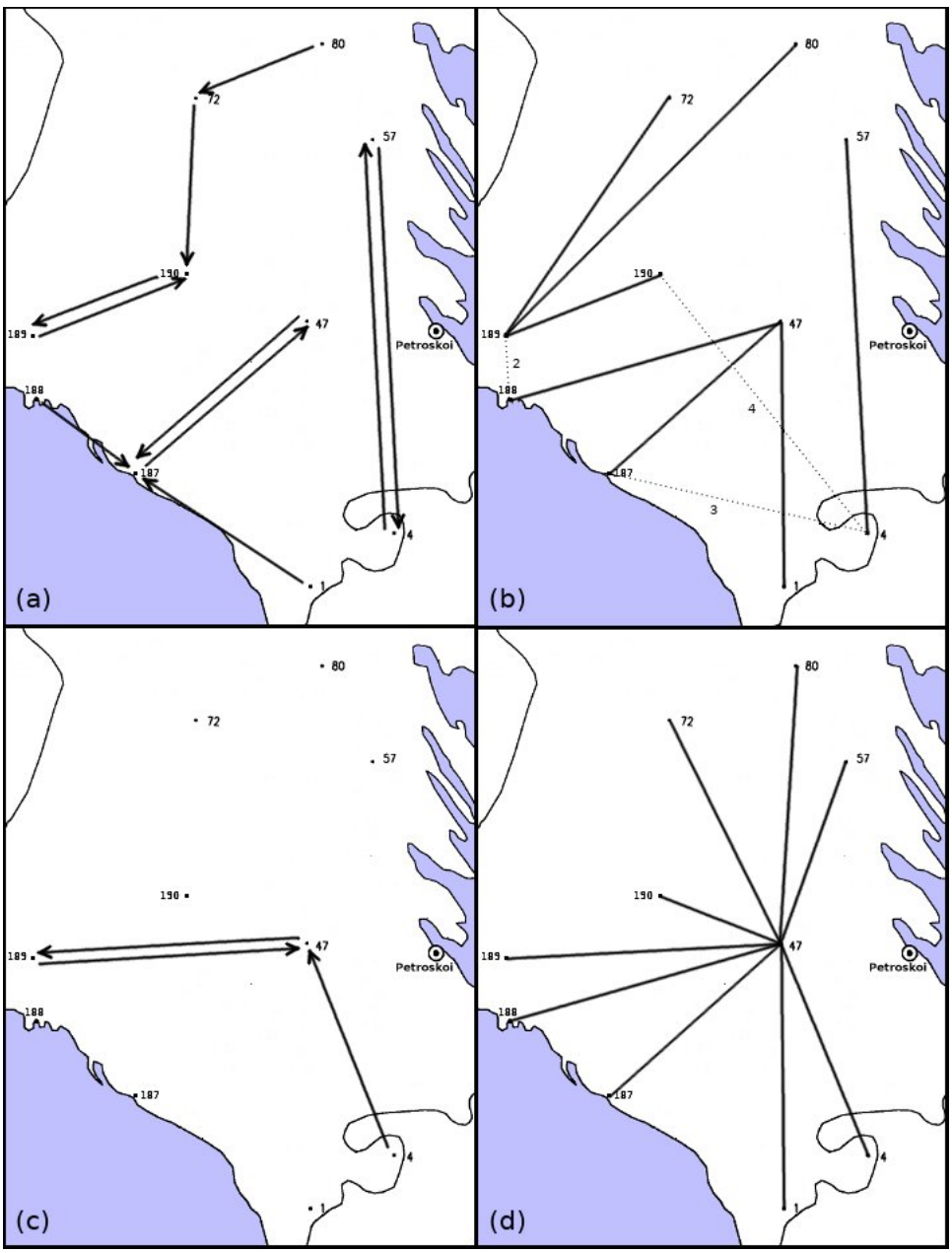

Figure 2. **Visualisation of the example. (a) Nearest neighbours. (b) Level 1 clusters. Dotted lines with numbers present differences of clusters. (c) Nearest neighbours of level 1 clusters. (d) Level 2 cluster(s).**

Note that on each level of clustering, the number of non-connected clusters decreases by half or more. Therefore, if the initial number of points is n, $\log_2(n)$ levels at most are needed to connect all points. As the Atlas with our extensions has 192 points, the number of levels never exceeds 7; in fact, depending on the selection of maps, it is usually much lower.

In order to answer the questions posed in the Introduction regarding the dialectal division of the Karelian language, and to demonstrate the

basic principles and capabilities of the methodology, in the following sections we analyse the cluster maps from the sections on phonetics and morphology, as well as the overview cluster of all Atlas materials.[2]

5. Analysis of sibilants on the cluster map

It is the phonetic system where Karelian dialects differ from each other most (Novak, Penttonen, Ruuskanen, Siilin 2019). Such a difference is, for example, the distribution of alveolar (*s*, *ś*, *z*, *ź*) and palato-alveolar (*š*, *ž*) sibilants (Figures 3—4). This complex phenomenon is studied in maps 74—99 of the Atlas.

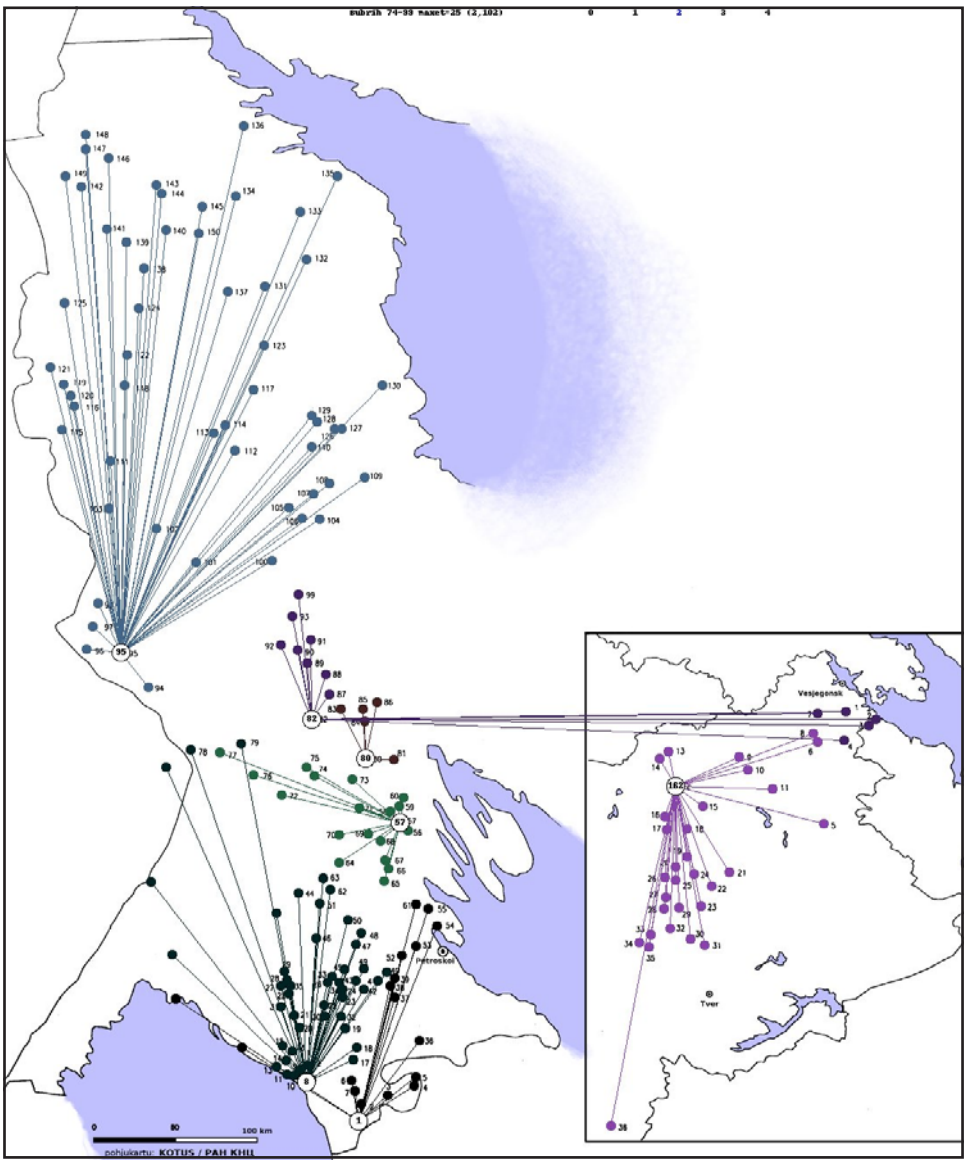

Figure 3. **The use of alveolar and palato-alveolar sibilants. Points 1—192, maps 74—99, level 3.**

[2] More analyses can be found in Novak, Penttonen, Ruuskanen, Siilin 2019 and cluster maps covering all Atlas sections on the web site http://karjalankieliopit.krc.karelia.ru.

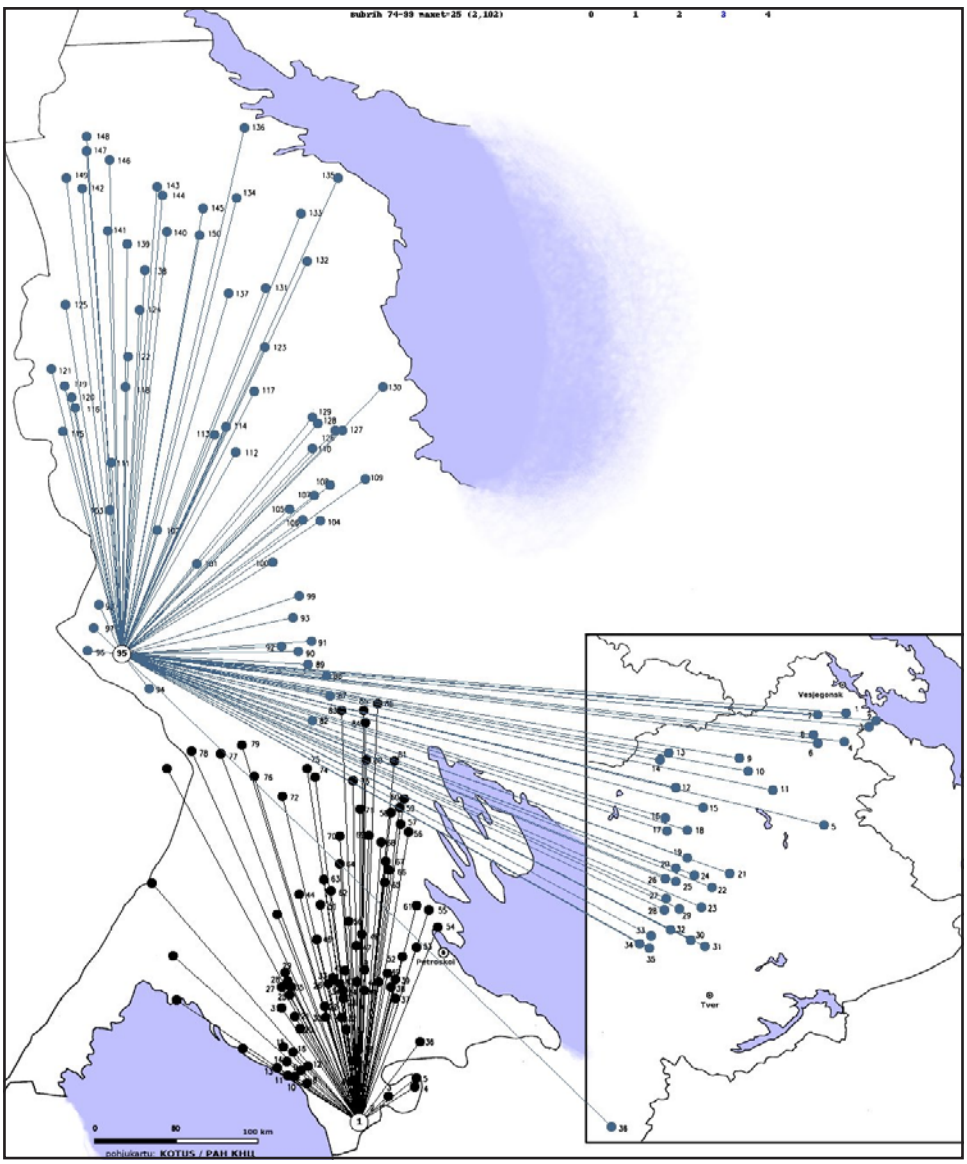

Figure 4. **The use of alveolar and palato-alveolar sibilants. Points 1—192, maps 74—99, level 2.**

In the cluster map on the distribution of alveolar and palato-alveolar sibilants, based on the Atlas maps 74—99, two large zones can be distinguished on level 3 (Figure 3): the northern and the southern one.

In the northern zone on level 2 (see Figure 4) three clusters are found:

- Cluster 95, consisting of the Viena Karelian subdialects, and South Karelian Reboľa, Rugarvi, and Tunkua subdialects;
- Cluster 162, consisting of Tver Karelian dialects;
- Cluster 82, consisting of the north-eastern subdialects of the dialect of Poodene and the eastern Vesjegonsk subdialects of Karelian Proper.

This area is characterized by the extensive use of palato-alveolar sibilants, while alveolar sibilants occur mainly after the vowel *i* (*šada* 'hundred',

šilmä 'eye', *istuu* 'sitting'). In cluster 82, alveolar sibilants are mainly used before front vowels (*šada*, *silmä*, *ištuu*).

In the southern zone, there are four clusters:

- Cluster 8 combines most of the Livvi Karelian subdialects (Kotkatjärvi, Vieljärvi, Videl, Tulemjärvi, Siämärvi), Border Karelian subdialects (Ilomantsi, Korbiselgä, Suistamo, Suojärvi) and some Poŕarvi subdialects of Karelian Proper;
- Cluster 1 contains the southern part of Ludic subdialects, the Nekkul and Riipuškal subdialects of Livvi Karelian, and the Border-Karelian Imbilahti and Salmi subdialects of Livvi Karelian;
- Cluster 57 consists of North and Central Ludic subdialects (excluding the southernmost ones), and the Mändyselgä and Por'arvi subdialects of Karelian Proper;
- Cluster 80 contains the subdialects of southwestern Poodene and the nearby Mändyselgä subdialects of Karelian Proper.

The southern zone as a whole is charaterized by an extensive use of alveolar sibilants, while palato-alveolar sibilants mostly occur after the vowel *i* (or *i**) (*sada, silmä, ištuu*). This distribution is inverted in the northern zone. Furthermore, cluster 57 differs from other clusters by using palato-alveolar sibilants even before the other front vowels (*šyvä*). In cluster 8, an alveolar sibilant can follow *i* (*istuu*), while in cluster 80, alveolar sibilants are used in all positions (*sada*, *silmä*, *istuu*).

In this case, the cluster map shows not only a phonetic phenomenon, but also illustrates the development of Karelian dialects. According to P. Virtaranta, palato-alveolar sibilants were not characteristic of the phonetic system of Old Karelian as all sibilants were alveolar (Virtaranta 1984 : 260). However, one cannot exclude the possibility that they appeared at the final stage of Old Karelian as a result of Russian influence, evidenced by the toponyms recorded in scribal books of the time, containing the sibilants *ж* (*ž*) and *ш* (*š*) (Kalima 1934 : 255—256). In modern Karelian dialects, the phenomenon appears in different ways. According to Bubrich, "once there were only two areas of alveolar and palato-alveolar sibilants — the southern and the northern one, and the border between them coincided with the border between the Livvi and Ludic dialect area and Karelian Proper" (Бубрих 1947b : 157—158). Obviously, in the process of further development around the border zone, as a result of influence from Livvi and Ludic dialects, the phonetic system of Poodene and Mändyselgä dialects, which later formed the transition zone, was significantly affected. The Tver Vesjegonsk subdialects of Karelian Proper, located at a distance of more than a thousand kilometers and more than three centuries away, are attracted to one of the clusters of this transitional zone. One explanation would be that immigrants, speakers of the same subdialect of the Old Karelian language, arrived in both territories. By another eplanation, Karelians from the territory of Poodene volost moved to Central Russia in the 18th century, according to a resident of the Poodene village of Selgi (Virtaranta 1961 : 42—44).

6. Analysis of local cases on a cluster map

We give another example of cluster maps concerning the Morphology section. The relative stability of the morphological system of a language against external influences explains the absence of a large number of significant

differences between the subdialects and dialects of the Karelian language at morphological level. But still there are some. These include, for example, the peculiarities of the dialectal paradigms of local cases, which are shown by the cluster map in Figure 5, compiled on the basis of Atlas maps 121—123. In this case, the analysis is based on a small amount of dialect material (three maps), which explains the appearance of fewer levels of clustering.

There are four clusters on this map:

- Cluster 72, comprising all Karelian Proper dialects, for which the syncretism of the external locative cases adessive and allative is characteristic (*mečässä* inessive singular 'in the forest', *mečästä* elative singular 'from the forest', *meččäh* illative singular, of the word *meččä* 'forest'; *veičellä* adessive singular, *veičeldä* ablative singular, *veičellä* allative singular from *veičči* knife);
- Cluster 6 consists of Livvi dialects (except for Nekkul and Riipuškal) and Border Karelian dialects, in which the inessive and elative cases (*mečäs*, *mečäs*, *meččäh*) as well as all external local cases (*veičel* : *veičel* : *veičel*) are identical;
- Cluster 1 consists of the Livvi dialects of Nekkul and Riipuškal, as well as South Ludic dialects, in which, unlike the previous group, the form of the allative is different from the rest of external local cases (*mečäs* : *mečäs* : *meččäh*; *veičel* : *veičel* : *veičele*);
- Cluster 4 includes Ludic dialects (except South Ludic), differing from cluster 1 in the illative case (*mečäs* : *mečäs* : *meččäi/meččähä*, *veičel* : *veičel* : *veičele*).

The reduction of local cases in the Livvi and Ludic dialects is attributed to Vepsian influence (Хямяляйнен 1961 : 92—109).

7. Analysis of Karelian dialects on the overview cluster map

From the point of view of linguistic geography, the dialect division of a language should take into account the isoglosses of various phonetic phenomena and morphological categories, as was the case presented above. In this section, we shall now analyse the dialect division from the point of view of a clusterization that is based on all Atlas maps 4—209, to get a general idea of the whole material (see Figures 6—7).

Based on maps 4—209, i.e. all dialect data of the Atlas, at level 3 of the clusterization, 3 clusters are found:

- Cluster 115 includes most of the dialects of Karelian Proper (except Poodene and Mändyselgä);
- Cluster 80 consists of the Poodene and Mändyselgä dialects of Karelian Proper, as well as the eastern Rugarvi dialects (Kuuziniemi, Korbilaksi). It is obvious that the formation of the cluster has been affected by the nearby Livvi and Ludic dialects. Thus, this cluster is located on the transitional zone between Karelian main dialects (transitional cluster);
- Cluster 16 comprising the Livvi and Ludic main dialects, as well as the dialects of Border Karelia, can be called the L i v v i-L u d i c cluster.

This result indicates that Livvi and Ludic dialects are closer to each other than to the dialects of Karelian Proper. Thus, based on the Atlas data, there are no linguistic grounds for considering Ludic as a distinct language. This map also shows that the dialects of Border Karelia have been signif-

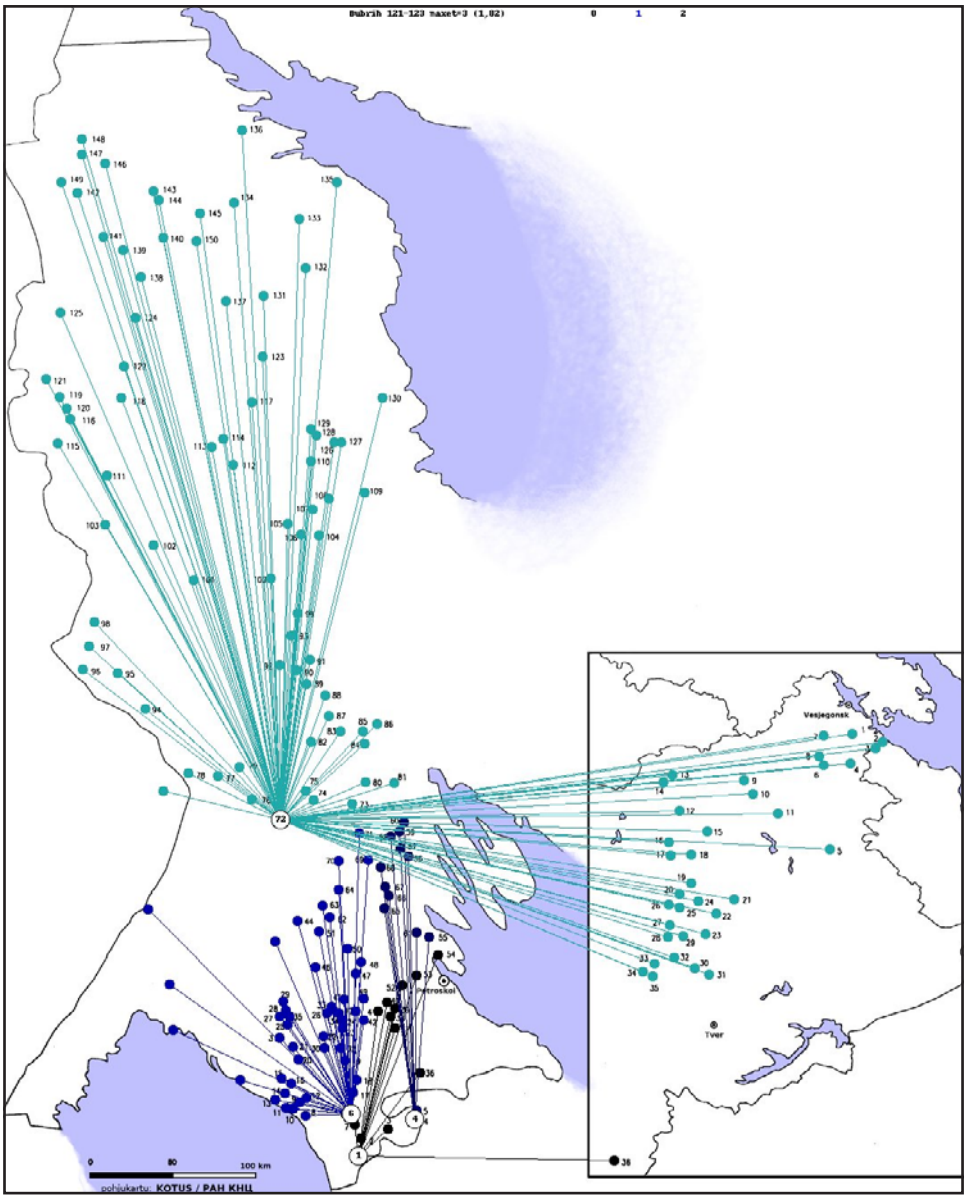

Figure 5. **Formation of local cases. Points 1—192, maps 121—123, level 1.**

icantly influenced by the neighbouring Livvi dialects, with which they have more features in common than with Karelian Proper.

Analysis of the cluster maps based on the thematic maps of the Atlas[3] allows us to draw the following conclusions:

Ludic dialects stand out as an independent main dialect with clear boundaries only as far as noun inflection is concerned. As for consonant and vocabulary systems, North and Central Ludic dialects bear a similarity to the southern Karelian dialects of Karelian Proper. Other Atlas sections show a

[3] Within the framework of the article, it is impossible to submit all thematic maps. You can familiarize yourself with them on the website http://karjalankieliopit.krc.karelia.ru or in Novak, Penttonen, Ruuskanen, Siilin 2019.

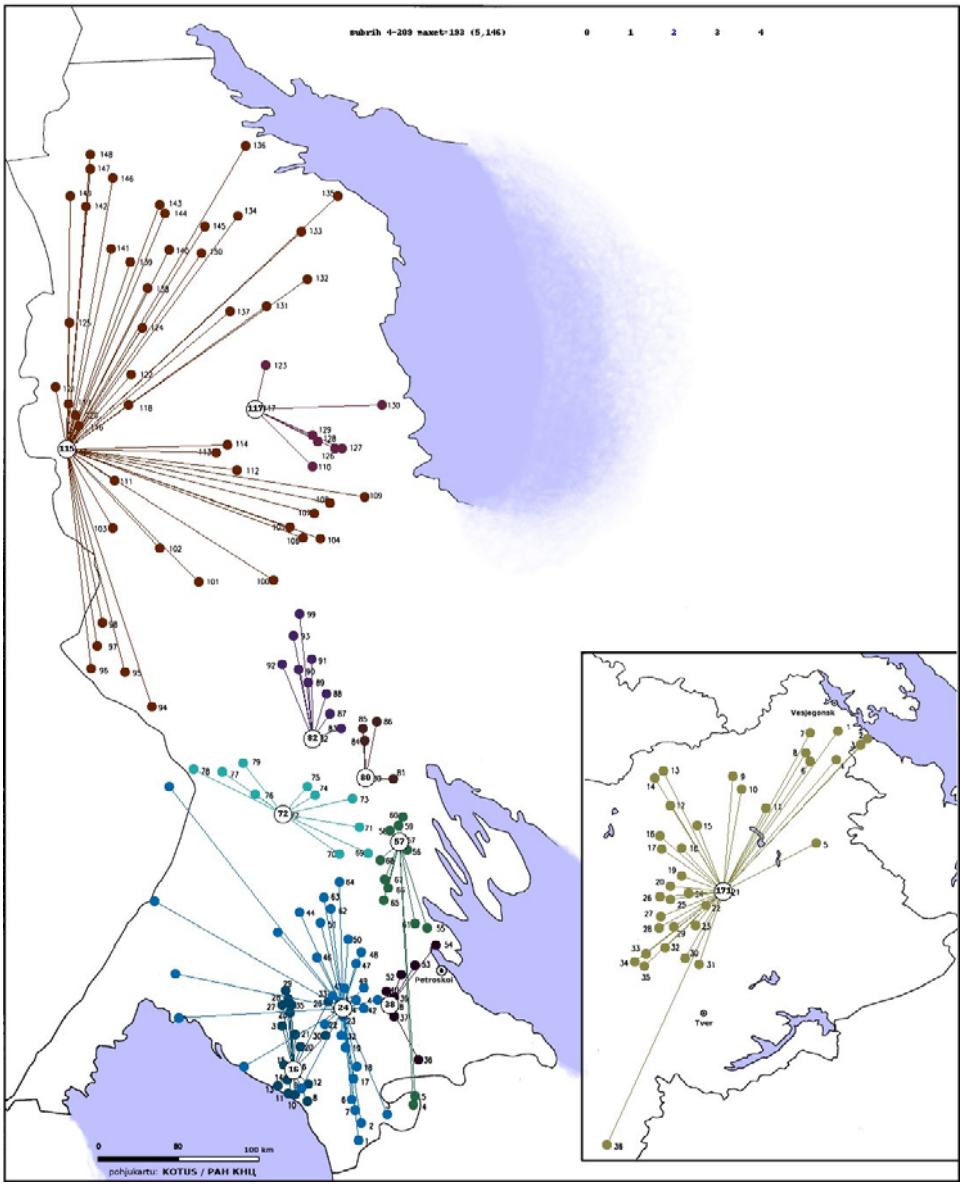

Figure 6. **Points 1—192, Atlas maps 4—209, level 3.**

similarity to Livvi dialects and consequently Ludic dialects belong to a common cluster with Livvi dialects.

Border Karelian subdialects belong to a common cluster with Livvi dialects in most of the Atlas sections. However, the vowel systems of Ilomantsi, Korbiselgä, Suistamo, and Suojärvi, as well as the verb and noun inflection systems of Ilomantsi, join them to Karelian Proper clusters;

The Tver dialects of the Karelian language form independent clusters separated from the rest of Karelian Proper on all maps, with the exception of the vocabulary map. The overview map on phonetics connects Tver dialects to the Northern and Central Karelian dialects. However, the same cannot be said of the Morphology map. Obviously, isolation for centuries far away

from other Karelian dialects and living in a Russian-speaking environment have left a significant imprint on the language of Tver Karelians. Yet, the eastern Vesjegonsk vowel system bears a similarity to central Karelian dialects, which suggests former migration from Central Karelia to that area.

7.1. The dialects of Karelian Proper on the overview cluster map

On level 2, cluster 115 of level 3 divides into 4 clusters, see Figure 7:

- Cluster 115 (on level 2) unites the Oulanka, Kiestinki, Kieretti, Viĉĉataipale, Pistojärvi, Uhut, Vuokkiniemi, Kontokki, Jyskyjärvi, Šuikujärvi, Tunkua, Rugarvi, and Rebol'a dialects;
- Cluster 117 in the east of this area includes the Paanajärvi, Užmana, and Voijärvi dialects;
- Cluster 72 covers the Poŕarvi and southern Mändyselgä dialects;
- Cluster 171 consists of Tver Karelian dialects.

An overview of the thematic maps shows that in terms of vowels, clusters 115 and 117 do not reveal significant differences, while cluster 72 even includes the Poodene and Border Karelian dialects (Suojärvi, Suistamo, Korbiselgä, and Ilomantsi dialects). Regarding consonants, Rugarvi and Rebol'a dialects deviate from cluster 115, forming a common group with the Voijärvi, Tunkua, Šuikujärvi, and Užmana dialects. This group of dialects also differs from the others in noun inflection. Regarding verb conjugation, cluster 115 splits into northern and southern zones. The southern zone includes the western part of cluster 72.

Cluster 117 groups in with Poodene dialects in verb conjugation. Thus, the separation of cluster 117 from cluster 115 is based on certain features of its conjugation system.

Cluster 72 shares verb conjugation with more northern dialects but otherwise differs from them, due to the location next to the Livvi Karelian dialects.

In thematic cluster maps, except for the vocabulary map, the Tver cluster 171 is divided into three zones — the core zone and two peripheral ones.

The transitional Poodene cluster 82 on level 3 is divided into two clusters on the second level. These two groups show differences in their vowel and consonant systems:

- Cluster 82 consists of the north-western part of level 3 cluster 82, which contains western Poodene and eastern Rugarvi dialects;
- Cluster 80 consists of the eastern Poodene and northern Mändyselgä dialects.

The data underlying the overview cluster map indicate that the division of Karelian Proper into North Karelian and South Karelian dialects is legitimate. However, the borderline between them needs closer examination, since the traditionally defined borderline (Figure 1) does not match with the border drawn by the clusterization. The group of Voijärvi, Tunkua, and Šuikujärvi dialects, as well as the Rugarvi and Rebol'a dialects here belong to the North Karelian group, although their consonant and noun systems form an independent cluster that is split into two regarding verb conjugation. These two clusters extend to the south and thus form a transitional group. At the same time, in terms of vowels, the group of Voijärvi, Tunkua, and Šuikujärvi dialects joins the dialects located north of it. In this way, the dialectal boundary coincides with the boundary between the former Arkhangelsk and Olonets gubernias.

The clusterization does not support the view that Jyskyjärvi, Paanajärvi, and Užmana dialects form a transitional zone between Viena and South Karelian dialects. Obviously, their common phonetic feature, the distribution of voiced and voiceless consonants, is not a sufficient ground for this.

7.2. Livvi and Ludic dialects on the overview cluster map

The Livvi-Ludic cluster 16 on level 3 of the overview map is split into 4 clusters on level 2, see Figure 7:

- Cluster 24 consists of Siämärvi, Vieljärvi, Kotkatjärvi, Nekkul, Riipuškal, as well as the Border Karelian Salmi and Impilahti subdialects of the Livvi dialect;
- Cluster 16 consists of the Videl and Tulemjärvi subdialects of Livvi Karelian;
- Cluster 38 consists of the Pyhärvi, Viidan and Kaškan subdialects of Ludic;
- Cluster 57 consists of the Kuďärv and northern (Hirvas, Koikari, Haldärv) subdialects of Ludic.

In terms of vowels, cluster 24 is divided into northern and southern zones. However, the consonant system and the verb conjugation system unite the eastern side of the Livvi cluster 24 with southern Ludic dialects (Pyhärvi, Viidan, Kuďärv). Partially this also holds for the vocabulary. At the same time, in terms of noun inflection a clear boundary is drawn between the Livvi and Ludic dialects.

8. Conclusion

Cluster analysis of the data collected in the Dialect atlas of the Karelian language indicates that there are two large dialect zones of the Karelian language: Karelian Proper and the Livvi-Ludic zone. There are a significant number of differences between them at all levels of language. At the same time, the Poodene and Tver dialects reveal more internal differences within the Karelian Proper dialect zone than the Ludic and Livvi dialects within the Livvi-Ludic zone.

In Karelian Proper, there are several large groups of dialects that reveal differences on all levels under consideration: phonetics, morphology, and vocabulary. However, the northern Viena Karelian group of dialects is relatively uniform in comparison to the South Karelian group, which has developed in the contact zone between the main dialects of Karelian.

According to our analysis, the group of dialects of Border Karelia (Ilomantsi, Suojärvi, Suistamo, and Korbiselgä), which developed in the border zone between the main dialects of the Karelian language, seem more like Livvi Karelian than Karelian Proper.

In contrast to the zone of Karelian Proper, the Livvi-Ludic dialect zone is more uniform, which is probably explained by the compactness of the territory. The analysis of the grammatical system of the dialects in this region makes it possible to separate Ludic from Livvi Karelian. Also, the grammatical system of Ludic subdialects, for example, the most southern and the most northern one, reveals fewer differences than, for example, the neighboring Videl and Nekkul Livvi subdialects.

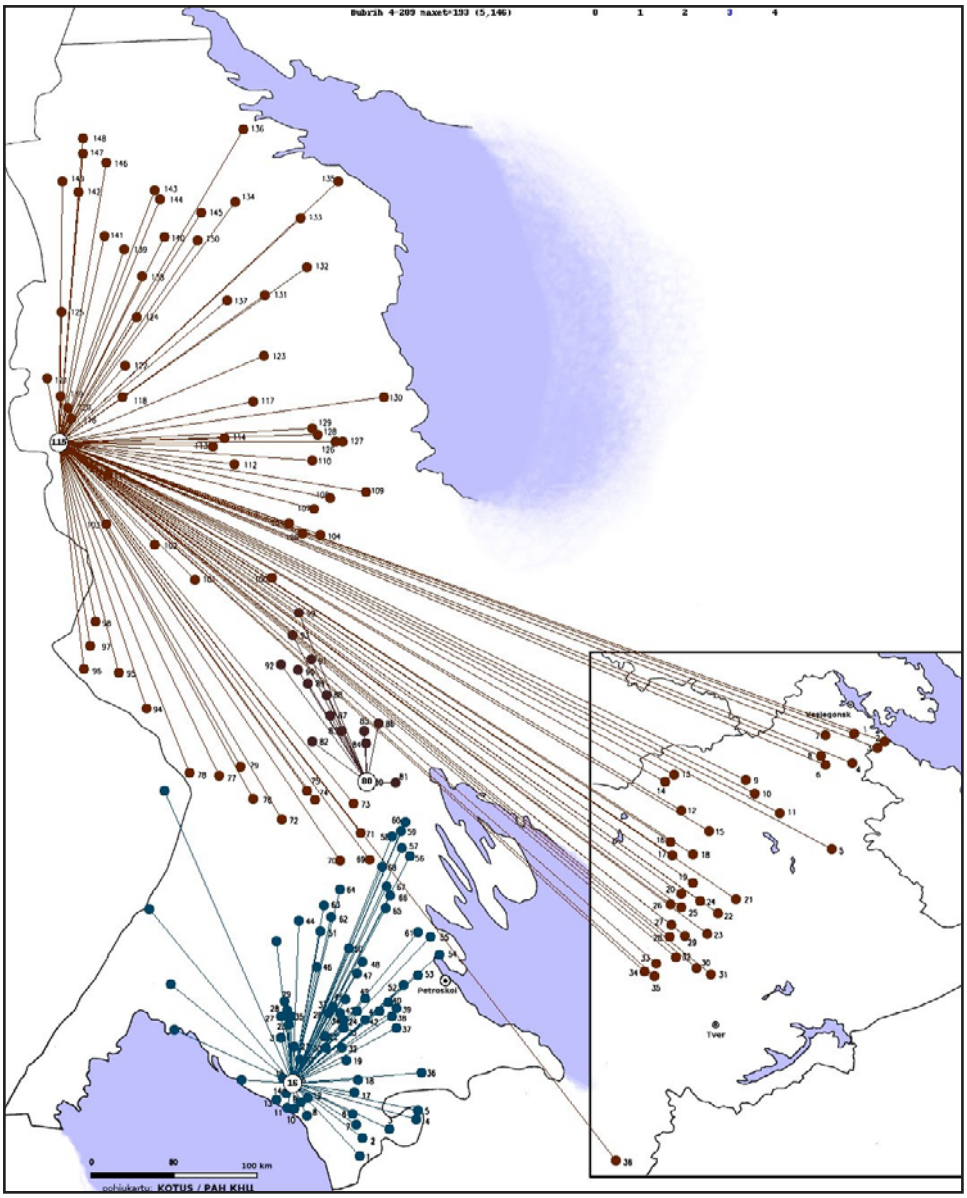

Figure 7. **Points 1—192, Atlas maps 4—209, level 2.**

A closer study of the answers to the questions in the data collection booklets (Бубрих 1937; 1946) could provide a more accurate understanding of the boundaries of the main dialects and dialects. Similar materials from Vepsian, Izhorian, and the eastern dialects of Finnish could further help to understand the differences between Karelian dialects.

We believe that clustering will provide a good basis for linguistic analysis to resolve the problem of the dialect division of the Karelian language. The data obtained at the first stage of our work clearly questions the validity of the traditional classification currently used by Russian and Finnish linguistics. The traditional classification and the classification obtained by us are based on the same linguistic features and represent the same period of

development of the Karelian language. The methodology for analyzing the available dialect data differs, as does the main principle of dialect division. Of course, divisions based on administrative division are understandable, but mostly it yields groups of subdialects with minor differences (usually phonetic ones), not a division of dialects.

The question of dialect division is conceptual: identification of distinct dialects assumes differences established on all levels of language: morphology, syntax, word formation, and vocabulary (JIEC 1990). In the traditional classification, in fact, subdialects and main dialects are described, but the division of dialects remains unsettled. The boundaries between separate groups of certain subdialects also need to be clarified, which can be done by applying the method described above. Of course, extralinguistic factors, such as the boundaries of political and socio-economic associations of different eras, ethnic self-awareness and the self-esteem of the people, elements of material and spiritual culture, mutual intelligibility between representatives of different groups of dialects, especially those on the border between two related languages (e.g. the Ludic dialects), must be taken into account. However, first and foremost, dialect division should be based on linguistic criteria.

It should be particularly noted that the dialect division discussed above and the corresponding map of the Karelian dialects can be considered relevant only for the mid-20th century, when most of the material for the Atlas was collected. At that time, the territory inhabited by Karelians was still relatively homogeneous in terms of linguistic composition, and the boundaries between subdialects and dialects were quite clear. At the present time, due to various historical and political events, the situation has changed significantly. Assimilative processes on both sides of the border have led to the disappearance of individual dialects, hence the boundaries of the living groups of dialects require clarification. The influence of the Russian language on the Karelian dialects, both in the Republic of Karelia and in Central Russia, has significantly increased. Border Karelian dialects have suffered from evacuation to Finland and experienced a significant influence from the Finnish side. Popularization of the standardized variants of the Karelian language also necessarily affects the dialectal speech of the Karelians. All of this means that the modern settlement map of the Karelians and, accordingly, the dialect map of the Karelian language differs from the traditional one.

The current language situation would require new research. The existing Atlas material is still useful for deciding on dialect divisions, for two reasons. Firstly, for a frozen dialect or spoken language, a period of a hundred years is not yet drastic, and secondly, gathering a collection of subdialect material of equal quality is no longer possible. The cluster analysis method presented above may also prove useful for solving problems that occur when rules and norms are developed for new variants of the written language. Finally, the method can prove useful for solving problems of dialect division for other languages.

Acknowledgements: We are grateful to Aleksi Ruuskanen and Lea Siilin for their comments and corrections to the manuscript. We are grateful to Maria Kok for her comments and for improving the language of the paper.

Addresses

Irina Novak
Institute of Linguistics, Literature and History of the Karelian Research Centre, Russian Academy of Sciences (Russia, Petrozavodsk)
E-mail: novak@krc.karelia.ru

Martti Penttonen
Department of Computer Science, University of Eastern Finland
E-mail: penttone@cs.uef.fi

Abbreviations

ALFE 1 — Atlas Linguarum Fennicarum. Itämerensuomalainen kielikartasto. Läänemeresoome keeleatlas. Ostseefinnischer Sprachatlas. Лингвистический атлас прибалтийско-финских языков, Helsinki 2004; **Atlas** — Д. В. Бубрих, А. А. Белякова, А. В. Пунжина, Диалектологический атлас карельского языка, Хельсинки 1997; **KKS** — Karjalan kielen sanakirja VI, Helsinki 2005 (LSFU XVI. Kotimaisten kielten tutkimuskeskuksen julkaisuja 25); **KKVS** — Karjalan kielen verkkosanakirja. https://kaino.kotus.fi/cgi-bin/kks/kks_etusivu.cgi; **ЛЕС** — Лингвистический энциклопедический словарь, Москва 1990. http://tapemark.narod.ru/les/.

LITERATURE

Borůvka, O. 1926, O jistém problému minimálním. — Práce Moravské přírodovědecké společnosti, sv. III, spis 3 1926, 37—58.
Genetz, A. 1872, Wepsän pohjoiset etujoukot, Helsinki.
—— 1880, Tutkimus Venäjän Karjalan kielestä. Kielennäytteitä, sanakirja ja kielioppi, Helsinki.
—— 1885, Tutkimus Aunuksen kielestä. Kielennäytteitä, sanakirja ja kielioppi, Helsinki.
Honkola, T., Santaharju, J., Syrjänen, K., Pajusalu, K. 2019, Clustering Lexical Variation of Finnic Languages Based on Atlas Linguarum Fennicarum. — LU LV, 161—184.
Itkonen, T. 1971, Aunuksen äänneopin erikoispiirteet ja aunukselaismurteiden synty. — Vir., 153—185.
Jájá, J. 1992, An Introduction to Parallel Algorithms, Reading, MA.
Jeskanen, M. 2019, Karjalan grammari kaikella rahvahalla 1, Tallinn.
Kalima, J. 1934, Entisen Käkisalmen läänin alueen aikaisemmasta kielimuodosta. — Vir., 254—256.
Koivisto, V. 2018, Border Karelian Dialects — a Diffuse Variety of Karelian. — On the Border of Language and Dialect, Helsinki, 56—84.
Lehtinen, J., Honkola, T., Korhonen, K., Syrjänen, K., Wahlberg, N., Vesakoski, O. 2014, Behind Family Trees. Secondary Connections in Uralic Language Networks. — Language Dynamics and Change 4, 189—221.
Leskinen, E. 1934, Karjalan kielen näytteitä II, Helsinki.
Leskinen, H. 1998, Karjala ja karjalaiset kielentutkimuksen näkökulmasta. — Karjala. Historia, kansa, kulttuuri, Helsinki, 352—382.
Novak, I., Penttonen, M., Ruuskanen, A., Siilin, L. 2019, Karjala kieliopeissa. Fonetiikan ja morfologian vertaileva tutkimus, Petroskoi (= Новак, Пенттонен, Руусканен, Сиилин 2019).
Ojansuu, H. 1907, Karjalan kielen opas. Kielennäytteitä, sanakirja ja äänneopillisia esimerkkejä, Helsinki.
Pahomov, M. 2017, Lyydiläiskysymys: Kansa vai heimo, kieli vai murre?, Helsinki.
Sammallahti, P. 1977, Suomalaisten esihistorian kysymyksiä. — Vir., 119—136.
Sollin, G. 1965, Le tracé de canalisation. — Programming, Games, and Transportation Networks, London—New York.
Turunen, A. 1946, Lyydiläismurteiden äännehistoria I. Konsonantit, Helsinki.

—— 1950, Lyydiläismurteiden äännehistoria II. Vokaalit, Helsinki.
Uusitupa, M., Koivisto, V., Palander, M. 2017, Raja-Karjalan murteet ja raja-alueiden kielimuotojen nimitykset. — Vir., 67—106.
Virtaranta, P. 1961, Tverin karjalaisten entistä elämää, Porvoo—Helsinki.
—— 1972, Die Dialekte des Karelischen. — СФУ 1, 7—27.
—— 1984, Über das *s* im Karelischen. — Studien zur phonologischen Beschreibung uralischer Sprachen, Budapest, 259—274.
Wiik, K. 2004, Karjalan kielen murteet. Kvantitatiivinen tutkimus. — FU 26, 239—302.
Беляков А. А. 1958, Языковые явления, определяющие границы диалектов и говоров карельского языка в КарАССР. — Прибалтийско-финское языкознание 12, 49—62.
Бубрих Д. В. 1937, Программа по собиранию материала для диалектологического атласа карельского языка, Петрозаводск.
—— 1946, Программа по собиранию материала для диалектологического атласа карельского языка, Петрозаводск.
—— 1947a, Происхождение карельского народа, Петрозаводск.
—— 1947b, Свистящие и шипящие согласные в карельских диалектах. — Ученые записки Ленинградского государственного университета 2, Ленинград, 129—159.
—— 1948, Историческое прошлое карельского народа в свете лингвистических данных. — Известия Карело-Финской базы Академии наук СССР 3, Петрозаводск, 42—50.
Бубрих Д. В., Беляков А. А., Пунжина А. В. 1997, Диалектологический атлас карельского языка, Хельсинки.
Зайков П. М. 1987, Диалектология карельского языка, Петрозаводск.
—— 1999, Грамматика карельского языка, Петрозаводск.
—— 2000, Глагол в карельском языке, Петрозаводск.
Керт Г. М. 2002, Карельская диалектология. К истории подготовки Диалектологического атласа карельского языка. — Очерки по карельскому языку, Петрозаводск, 20—38.
Керт Г. М., Рягоев В. Д. 1997, Предисловие к «Диалектологическому атласу карельского языка». — Диалектологический атлас карельского языка, Хельсинки, 1—4.
Новак И., Пенттонен М., Руусканен А., Сиилин Л. 2019, Карельский язык в грамматиках. Сравнительное исследование фонетической и морфологической систем, Петрозаводск (= Novak, Penttonen, Ruuskanen, Siilin 2019.)
Рягоев В. Д. 1993, Карельский язык. — Языки мира. Уральские языки, Москва, 63—75.
Хямяляйнен М. М. 1961, О развитии внутренне-местных падежей в северо-восточной группе прибалтийско-финских языков. — Прибалтийско-финское языкознание, Москва—Ленинград, 84—109.

ИРИНА НОВАК (Петрозаводск), *МАРТТИ ПЕНТТОНЕН* (Куопио)

ИСПОЛЬЗОВАНИЕ АЛГОРИТМА КЛАСТЕРНОГО АНАЛИЗА В РЕШЕНИИ ВОПРОСОВ ДИАЛЕКТНОГО ЧЛЕНЕНИЯ КАРЕЛЬСКОГО ЯЗЫКА

В статье рассматриваются возможности применения агломеративно-иерархического метода кластерного анализа к материалам «Диалектологического атласа карельского языка» (Бубрих, Беляков, Пунжина 1997), что позволяет наметить пути решения проблем карельской диалектологии. Кластеризация может производиться на базе, как всех материалов Атласа, так и отдельных его тематических разделов, например, морфология, именная словоизменительная

система или лексика. Методика рассматривается на примере анализа дистрибуции переднеязычных щелевых согласных, местных падежей, а также всех материалов Атласа в совокупности. Результаты кластеризации, подтверждающиеся исследованиями по карельской диалектологии, позволяют сделать следующие выводы. Диалекты карельского языка представляют собой две крупные зоны: собственно карельскую и ливвиковско-людиковскую. Говоры Приграничной Карелии обнаруживают большее число общих черт с ливвиковскими диалектами, чем с собственно карельскими. Нет лингвистических оснований для выделения столь большого числа диалектов, как это произведено в основанной на административном принципе традиционной классификации карельского языка.

IRINA NOVAK (Petroskoi), *MARTTI PENTTONEN* (Kuopio)

KLASTERANALÜÜSIL PÕHINEV KARJALA MURDEJAOTUS

Autorid tutvustavad algoritmi, mida on võimalik kasutada karjala murdejaotuse täpsustamiseks murdeatlases (Бубрих, Беляков, Пунжина 1997) esitatud keeleandmete sarnasuse alusel. Artiklis on vaadeldud klastreid, mis põhinevad sibilantidel, kohakäänetel ja kogu atlasel. Klastrite analüüs võimaldab teha esiteks järelduse —mida toetab ka senine uurimistöö —, et karjala murded võib jagada kaheks: päriskarjala ning livvi (aunuse) ja lüüdi keelalaks. Teiseks, et Soome piiri äärsetel murretel on rohkem ühisjooni livvi kui päriskarjala keelega. Kolmandaks ilmneb, et traditsiooniline valdadel põhinev eristus ei ole keeleliselt küllaldaselt põhjendatav.