УДК 811.511.112’28’342.415(470.22):004.93’14
DOI: 10.30624/2220-4156-2022-12-3-486-496

Система восходящих дифтонгов в говорах карельского языка Карелии: сравнение методов кластеризации

И. П. Новак
Институт языка, литературы и истории
Карельского научного центра Российской академии наук,
г. Петрозаводск, Российская Федерация,
novak@krc.karelia.ru

Н. Б. Крижановская
Институт прикладных математических исследований
Карельского научного центра Российской академии наук,
г. Петрозаводск, Российская Федерация,
nataly@krc.karelia.ru

АННОТАЦИЯ

Введение. В последнее десятилетие в финно-угроведении набирают популярность статистические методы диалектологии. Результаты первого этапа применения методики кластеризации к материалам «Диалектологического атласа карельского языка» (1997) проявили основные проблемы карельской диалектологии (несостоятельность традиционной классификации, нечёткое определение статуса и границ отдельных групп говоров и пр.). В целях их решения была создана современная диалектная база данных карельского языка, включающая закодированные языковые данные, что сделало возможным применение к ним различных иерархических и итеративных методов кластеризации.

Цель: определение метрики для верификации и уточнения существующей схемы диалектного членения карельского языка на примере анализа системы восходящих дифтонгов.

Материалы исследования: оцифрованные и закодированные данные «Программ по собиранию материала для диалектологического атласа карельского языка», заполненные в 1937–1972 гг.

Результаты и научная новизна. Научная новизна заключается в применении к большим объёмам карельского диалектного материала статистических методов диалектометрии. В рамках исследования проведено пять видов кластеризации, демонстрирующих распределение вариантов восходящих дифтонгов в карельских говорах Карелии: методом полной связи (три кластеризации), центроидным иерархическим методом и методом k-средних. Результаты кластеризаций не обнаруживают существенных отличий между собой, но наилучшим образом (при сравнении визуализированных данных вручную) проявили себя методы полной связи и k-средних. Итоговая кластерная карта в целом совпала с картиной, описанной в исследованиях по карельской фонетике и диалектологии, но позволила получить более чёткое представление о границах анализируемого диалектного явления и его переходных зонах. Это доказывает правомерность применения методики для решения проблем карельской диалектологии, а также в процессе переработки диалектной классификации языка.

Ключевые слова: диалектология, лингвистическая география, диалектометрия, кластерный анализ, метод кластеризации, карельский язык, восходящие дифтонги

Благодарности: Работа И. П. Новак выполнена в рамках бюджетного финансирования КарНЦ РАН (№ 121070700122-5); Н. Б. Крижановской – за счёт гранта Российского научного фонда № 22-28-20215 «Создание речевого корпуса прибалтийско-финских языков Карелии», проводимого совместно с органами власти Республики Карелия с финансированием из Фонда венчурных инвестиций Республики Карелия (ФВИ РК).

Для цитирования: Новак И. П., Крижановская Н. Б. Система восходящих дифтонгов в говорах карельского языка Карелии: сравнение методов кластеризации // Вестник угроведения. 2022. Т. 12. № 3. С. 486–496.

The system of ascending diphthongs in dialects of the Karelian language: comparison of clustering methods

I. P. Novak
Institute of Linguistics, Literature and History,
Karelian Research Centre of the Russian Academy of Sciences,
Petrozavodsk, Russian Federation,
novak@krc.karelia.ru

N. B. Krizhanovskaya
Institute of Applied Mathematical Research,
Karelian Research Centre of the Russian Academy of Sciences,
Petrozavodsk, Russian Federation,
nataly@krc.karelia.ru

ABSTRACT

Introduction: in the last decade, statistical methods of dialectology are increasingly used in Finno-Ugric studies. The results of the first stage of applying the clustering technique to the materials of the Dialectological Atlas of the Karelian Language (1997) revealed the main problems of Karelian dialectology (failure of the traditional classification, unclear definition of the status and boundaries of certain groups of dialects, etc.). To solve these problems, a dialect base of the Karelian language was developed. This base includes encoded language data, which made it possible to apply various hierarchical and iterative clustering methods to this data. This base includes encoded language data to which various hierarchical and iterative clustering methods are applied.

Objective: choice of a metric and a clustering method for verification and refinement of the existing scheme of the existing scheme of dialect division of the Karelian language, on the example of the analysis of the system of ascending diphthongs.

Research materials: digitized and coded data of the "Programs for collecting material for the dialectological atlas of the Karelian language", completed in 1937–1972.

Results and novelty of the research: scientific novelty is the application of statistical methods of dialectometry to large volumes of Karelian dialect material. During the study, five variants of clusterization were carried out, demonstrating the distribution of variants of ascending diphthongs in the Karelian dialects of Karelia: the complete-linkage method (three clusterizations), the centroid linkage method, and the k-means method. The results of clusterizations do not show significant differences, but the methods of complete-linkage and k-means showed themselves in the best way. The final cluster map coincided with the picture described in studies on Karelian phonetics and dialectology, but made it possible to obtain clearer boundaries of the analyzed dialect phenomenon and its transitional zones. This proves the legitimacy of applying the methodology for solving the problems of Karelian dialectology, as well as in the process of reworking the dialect classification of the language.

***Key words*:** dialectology, linguistic geography, dialectometry, cluster analysis, clustering method, Karelian language, ascending diphthongs

Acknowledgements: the study was carried out under the state order of the Karelian Research Centre of the Russian Academy of Sciences (№ 121070700122-5) and through Russian Science Foundation grant 22-28-20215 Creation of the speech corpus of the Baltic-Finnic languages of Karelia implemented in collaboration with Republic of Karelia authorities with funding from the Republic of Karelia Venture Capital Fund.

For citation: Novak I. P., Krizhanovskaya N. B. The system of ascending diphthongs in dialects of the Karelian language: comparison of clustering methods // Vestnik ugrovedenia = Bulletin of Ugric Studies. 2022; 12 (3): 486–496.

Введение

Развитие статистических методов диалектологии, продиктованное накоплением больших объёмов диалектных данных и сложностями их анализа, относится к началу 70-х гг. XX столетия [12; 19]. К настоящему моменту многие европейские языки уже стали объектом диалектометрии [13; 22], в последнее десятилетие она набирает популярность и в финно-угроведении [11; 1], поскольку позволяет в краткие сроки обрабатывать объёмные массивы данных и решать такие задачи лингвистической географии, как проведение границ между диалектами с целью переработки или уточнения диалектных классификаций языков. Для диалектометрических исследований используются диалектологические атласы, диалектные базы данных или материалы языковых корпусов.

Методика использования кластерного анализа (группировка множества объектов по целому набору признаков) применительно к карельскому диалектному материалу в 2018–2021 гг. была апробирована на базе материалов «Диалектологического атласа карельского языка» [3] и «Сопоставительно-ономасиологического словаря диалектов карельского, вепсского и саамского языков» [9] с использованием метода кластеризации Соллина [20; 7]. Результаты первого этапа работы позволили выявить основные проблемы карельской диалектологии (несостоятельность традиционной классификации, проблема определения статуса отдельных групп говоров, проблема терминологии, проблема выбора основного принципа диалектного членения) и сделать вывод о возможности успешного применения статистических методов для их решения.

В качестве объекта анализа выбран один из ярчайших фонетических маркеров карельской диалектной речи – система восходящих дифтонгов

(вторым компонентом выступают гласные *o, ö, e, a, ä*). Для карельского языка, в отличие от остальных прибалтийско-финских языков, характерна обильная дифтонгизация, восходящая к древнекарельскому периоду её развития. Прибалтийско-финские праязыковые долгие гласные **oo* > *uo*, **öö* > *yö*, **ee* > *ie* подверглись дифтонгизации первыми, поскольку данное явление характерно для всех прибалтийско-финских языков [8, 42, 45; 18, 370–371, 379; 21, 355; 23, 20, 25, 33;]. Начало процесса дифтонгизации долгих **aa* > *oa* > *ua*, **ää* > *eä* > *iä* финно-угроведы относят к XI–XII вв., т. е. к тому периоду, когда ижорский язык уже отделился от древнекарельского, на что указывает сохранение в нём долгих гласных, а восточный диалект финского языка Саво, обнаруживающий дифтонги, ещё не успел оформиться в самостоятельный [21, 350–352; 16].

Прибалтийско-финские праязыковые долгие гласные ударных слогов, как и долгие гласные или сочетания гласных, возникшие в более поздний период вследствие стяжения, представлены в современных карельских говорах следующим набором рефлексов: **aa* > *oa* / *ua* / *ja* / *uo* / *oo* / *aa*; **ää* > *eä* / *iä* / *ie* / *ee* / *ää* / *öä*. Их распределение в отдельных диалектах / группах говоров карельского языка исследователи карельской фонетики и диалектологии описывали с конца XIX в. [2, 45–46; 10, 33–43; 14, 12; 15, 165–166; 23, 17–19, 31; 24, 272; 25, 17–18]. В «Диалектологическом атласе карельского языка» (ДАКЯ) содержатся две карты с представительством прибалтийско-финских долгих гласных *aa*, *ää* первого слога, и 2 карты, демонстрирующие итог стяжения гласных при выпадении интервокальных согласных (*a-a, ä-ä*). Их ручное сопоставление с привлечением данных дескриптивных описаний отдельных диалектов позволяет говорить о шести основных диалектных ареалах анализируемого явления (рис. 1.):

1) собственно карельский (без южных ругозерских, паданских, поросозерских и южных мяндусельгских говоров) – *ua*, *iä* (в ударном слоге), *ua*, *yä*/*iä*/*öä* (вследствие стяжения);

2) ливвиковско-южнокарельский (поросозерские, южные мяндусельгские, отдельные паданские, сямозерские, ведлозерские и коткозерские говоры) – *oa*, *eä*;

3) собственно карельский паданский – *oo*, *ee*;

4) ливвиковский ведлозерско-тулмозерский – *aa*, *ää*/*ee*;

5) ливвиковский неккульский (+ среднелюдиковские говоры) – *uo*, *ie*;

6) людиковский – *ua*, *iä* (в ударном слоге), *ada*, *ädä* (отсутствие стяжения) [3, к. 4–7].

Рис. 1. Схема диалектных ареалов представительства восходящих дифтонгов в говорах карельского языка

Целью настоящего исследования является определение наиболее подходящего для решения проблем карельской диалектологии алгоритма кластеризации на примере анализа системы восходящих дифтонгов карельского языка.

Материалы и методы

В целях решения проблем карельской диалектологии была разработана[1] диалектная база карельского языка Murreh [5]. В неё вошли предварительно закодированные архивные данные заполненных лингвистами в 1937–1972 гг. «Программ по собиранию материала для диалектологического атласа карельского языка» (более 200 тыс. ед.)[2]. Эти материалы являются пригодными для решения

[1] Комплекс программ разработан Н. Б. Крижановской.

проблем диалектного членения карельского языка в силу определённой устойчивости его диалектной речи.

Диалектный материал по анализируемому явлению (рефлексы **aa*, **ää*, **a-a*, **ä-ä*) в базе данных представлен 18 вопросами и ответами на них, полученными в 146 населённых пунктах (говорах) Карелии[3], т. е. речь идёт о более чем 2,5 тыс. единиц для проведения кластеризации.

На подготовительном этапе исследования строится матрица расстояний между объектами (говорами). При этом за **метрику** (расстояние между двумя объектами) принимается количество различных ответов на одинаковые вопросы, отсутствие ответа обсчитывается как ½ единицы.

Модуль кластеризации диалектной базы карельского языка позволяет применять к языковым данным как иерархические, так и итеративные статистические методы анализа.

В **иерархических методах** объекты шаг за шагом объединяются в более крупные кластеры. Эти методы отличаются способом определения «расстояния» между кластерами и стратегией выбора кластеров для объединения (рис. 2). Так, например, в **методе одиночной связи** расстояние между двумя кластерами равно расстоянию между их ближайшими представителями, в **методе полной связи**, наоборот, – между самыми дальними объектами, а в **центроидном методе** – между центральными объектами кластеров [6, 114–115; 4, 16–21].

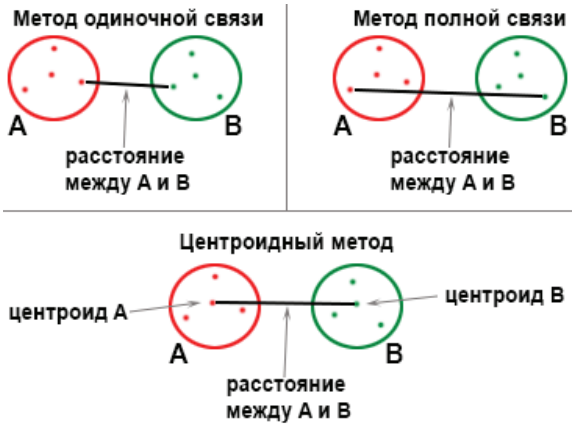


Рис. 2. Иерархические методы кластеризации

Для иерархических методов характерна такая особенность, что если один элемент попал в какой-либо кластер, то на следующих шагах он из этого кластера уже выйти не может. Наличие возможности перехода элемента из кластера в кластер предоставляет **метод k-средних (итеративный метод)** [6, 115; 4, 21–22].

Результаты кластеризации программа визуализирует в виде карт, что существенно облегчает задачу их анализа.

Результаты

1. Кластеризация диалектного материала методом полной связи

В результате применения метода полной связи[4] говоры[5] оказались разбиты на 5 относительно крупных (13–52 н. п.) и 15 более мелких (1–6 н. п.) кластеров (рис. 3а):

– крупный **красный кластер** говоров (52), выделившийся на севере территории (собственно карельские говоры) характеризуется преимущественным употреблением дифтонгов *ua*, *iä* в качестве рефлексов праязыковых **aa*, **ää* ударного слога слова: *mua* ‘земля’[6], *ruato* ‘работа’, *huapat* ‘осины’, *muamo* ‘мать’, *kuatiet* ‘порты’, *šuat* ‘до’; *piä* ‘голова’, *viärä* ‘кривой’, *piästäy* ‘пускать’, *piäšöy* ‘попадать’; в то время как в результате стяжения гласных образовались дифтонги *ua*, *yä*: *muattih* ‘они спали’, *oštua* ‘покупать’, *harmuat* ‘серые’, *vanhua* ‘старого’, *kaivua* ‘копать’; *elyä* ‘жить’, *tietyä* ‘знать’, *kyntyä* ‘пахать’. К кластеру примыкают отдельные ливвиковские сямозерские и ведлозерские говоры;

– **оливковый кластер** (13) объединил собственно карельские юго-восточные паданские, северные мяндусельгские и ребольские говоры, говоры д. Каменное Озеро (115) и Тунгуда (107), а также отдельные ливвиковские говоры, в которых дифтонги *ua*, *iä* используются во всех анализируемых позициях: *mua, ruado, huavat, muamo, kuadiet, suat; piä, viärä, piästav, piäzöv; muattih, ostua, harmuat, vahnua, kaivua; eliä, tiediä, kyndiä*. При дальнейшей кластеризации оливковый и красный кластеры объединяются;

---

[2] Программы по собиранию материала для диалектологического атласа карельского языка // Научный архив КарНЦ РАН: Ф. 1. Оп. 38. Д. 16–170, 261–263, 267, 268, 276; Оп. 43, № 43–74, 116–132, 138, 140, 179–181, 195, 219–225, 274, 275. Петрозаводск, 1937–1950.

[3] Познакомиться с полным списком вопросов и вариантами ответов можно на сайте базы в соответствующем разделе: http://murreh.krc.karelia.ru/ques/question

[4] Познакомиться с интерактивной кластерной картой и дендограммой кластеризации можно на сайте базы: http://murreh.krc.karelia.ru/experiments/anketa_cluster/example/1_2

[5] Под термином «говор» подразумевается одна единица кластеризации (один населенный пункт).

[6] Перевод примеров даётся один раз при первом упоминании. Для крупных кластеров приводятся лишь наиболее распространённые фонетические варианты ответов.

– **малиновый кластер** (16), в который вошли преимущественно людиковские говоры, также характеризуется использованием дифтонгов *ua*, *iä* в качестве рефлексов праязыковых ударных **aa*, **ää*: *mua, ruad, hoabad, muamo, kuad'd'at, suat; piä, viär, piästäv, piäzöv.* Характерной чертой людиковского наречия является отсутствие чередования согласных, в связи с чем на месте прибалтийско-финских интервокальных спирантов, выпавших в других наречиях, что привело к стяжению гласных в дифтонги, здесь представлены звонкие смычно-взрывные согласные, а, соответственно, и менее обильная дифтонгизация: *magattih, ostada, harmagat, vanhad, kaivada; elädä*;

– для крупного **небесно-голубого кластера** (23), объединившего собственно карельские паданские, поросозерские и мяндусельгские, а также ливвиковские сямозерские, ведлозерские и некоторые коткозерские говоры характерно доминирующее употребление во всех позициях дифтонгов *oa*, *eä*: *moa, roado, hoavat, moamo, koadiet, soat; peä, veärä, peästäv, peäzöv; moattih, ostoa, harmoat, vahnoa, kaivoa; eleä, tiedeä, kyndeä*;

– незначительные отличия от предыдущего обнаруживает **коричневый кластер** (6), представленный в основном ливвиковскими сямозерскими говорами, в которых преобладает употребление дифтонгов *oa*, *eä*, но наряду с ними встречаются также случаи использования дифтонгов *ua*, *iä*: *moa / mua, roado, hoavat, koadiet, soat; peä / piä, veäry / viäry, peästäv, peäzöv; moattih / muattih, ostua / ostoa, harmoat / harmuat, vahnua, kaivoa / kaivua; eliä / eleä, tiediä, kyndiä*. На следующих уровнях кластер объединяется с крупным небесно-голубым;

– **зелёный кластер** (10), представлен ливвиковскими видлицкими и отдельными тулмозерскими говорами, где во всех анализируемых позициях выступают преимущественно долгие гласные *aa*, *ää*: *maa, raado, haavat, maamo, kaadiet, saat; pää, vääry, päästäy, pääzöv; maattih, ostaa, harmaat, vanhaaa, kaivaa; elää, tiedää, kyndää*;

– отдельного внимания заслуживают соседствующие **розовый кластер** (3), представленный собственно карельскими паданскими говорами д. Паданы (87), Сяргозеро (89) и Сондалы (88), и **бирюзовый кластер** (2), включивший говоры д. Лазарево (92) и Гимолы (79). Для них характерно использование долгих гласных, выступающих в кластерах в разных пропорциях: *moo / maa, roodo, hoobat / haabat, moomo, koodiat; pee, veerä, peestee, peezöö; moottih, ostoo, harmoot, vanhoo / vanhaa, kaivoo / kaivaa; elee / elää, tiedee / tiedää, kyndee / kyndää*;

– **голубой кластер** (2), представленный говорами д. Куйтежа (2) и Мегрозеро (3), выделяется на фоне основной массы ливвиковских говоров возможностью использования дифтонгов *uo*, *ie.* Непосредственно к нему при дальнейшей кластеризации примыкает и **пурпурный** кластер (2), в который вошли говоры д. Мегрега (1) и Габозеро (24): *muo, ruodo, huovat, muomo, kuadiet, suat; piä, viery, piestav, piezöu; muottih, ostua, harmuat, vahnua, kaivua, kaivuo; elie, tiedie, kyndie*. Возможность употребления дифтонгов *uo*, *ie* отличает также **коралловый кластер** (2), в который вошли людиковские говоры д. Гомсельга (61) и Намоево (55), отличающийся, однако, от предыдущих двух менее обильной дифтонгизацией: *magattih, ostada, harmagat, vahnad, kaivada; elädä, tieta, kyntä*.

Оставшиеся кластеры представлены одним–тремя говорами. Их диалектные данные указывают на смешение в них различных вариантов анализируемого явления в самых разных пропорциях. На следующих шагах эти кластеры либо присоединяются к соседним, либо на протяжении большого числа шагов продолжают организовывать самостоятельные кластеры, что может указывать на значительный процент смешения в них различных фонетических вариантов рассматриваемого явления.

На карте (рис. 3а) бросаются в глаза отдельные «выбросы» (пункты, не находящиеся вблизи основных диалектных границ анализируемого явления, но отличающиеся по цвету от окружения), сохраняющиеся на протяжении довольно большого количества шагов кластеризации (например, Парфеево (134) или Войница (121)), что может являться следствием дефицита материала по данным говорам (Войница) или служить сигналом возможного наличия ошибок в привлечённых диалектных данных (Парфеево: сбор осуществлялся карелом-ливвиком, носителем коткозерского говора, это могло сказаться на результате). Процент подобных «выбросов» невысок, что позволяет использовать метод для решения поставленных задач.

Применение дополнительных параметров кластеризации позволяет получить несколько отличные результаты. Так, например, для определения принадлежности к тому или иному кластеру говоров, обнаруживающих дефицит материала, можно отключить подсчёт отсутствия

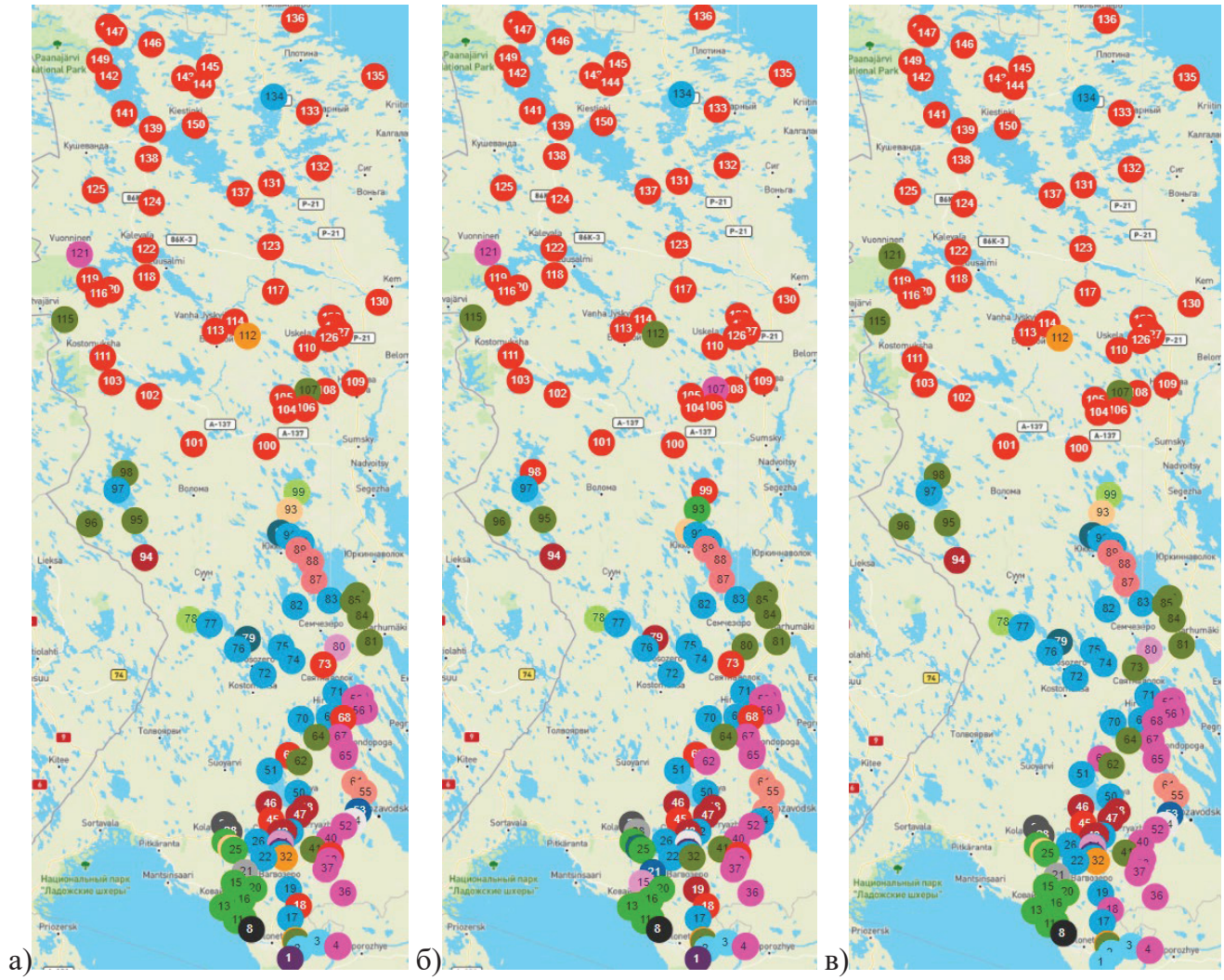


Рис. 3 Кластеризация методом полной связи: а) по умолчанию;
б) без учёта возможного отсутствия некоторых ответов; в) с учётом географического положения

ответа в анкетах как отличия (0,5). Изменение расчёта метрики (отсутствие ответа считается совпадением) позволяет обнаружить незначительные отличия (см. рис. 3б): перераспределилось лишь несколько говоров (главным образом, расположенных на периферии крупных кластеров), что объясняется смешением в них различных вариантов исследуемого явления. В случае дефицита материала параметр “НЕ считать отсутствие ответа как отличие” может как помочь разобраться с проблемными моментами, так и усложнить картину, поэтому в таких случаях требуется индивидуальная работа по каждому из “гуляющих” пунктов.

Метод полной связи предполагает, что на каждом следующем шаге для сравнения двух кластеров выбирается по одному говору из каждого кластера, между которыми сравниваются расстояния. В процессе отбора автоматически действует алфавитный порядок определения пунктов для сравнения. В таком случае, как показали результаты рабочих экспериментов, в один кластер могут попасть говоры, территориально находящиеся друг от друга на довольно большом расстоянии, при этом соседние говоры могут большее количество шагов оставаться в разных кластерах. В связи с этим в программу был включён дополнительный параметр, позволяющий учитывать географическое положение в процессе отбора пунктов для сравнения, поскольку велика вероятность того, что говоры, находящиеся рядом, наиболее близки друг к другу. Учёт географического положения говоров (рис. 3в) при обработке тех же диалектных материалов позволил обнаружить, опять же, незначительные отличия в результатах кластеризации. В этот раз изменение затронуло ещё меньшее число говоров, что, однако, помогло справиться с некоторыми «выбросами» (Уссуна (68), Пелдожа (39), Войница (121)). При этом основные переходы произошли между близкими кластерами, а часть изменений оказалась связана с нехваткой исходного материала.

2. Кластеризация центроидным иерархическим методом

При разбиении говоров на 20 кластеров центроидным иерархическим методом (рис. 4а):

– на севере территории выделился крупный «собственно карельский» кластер (красный, 71 пункт), объединивший близкие друг к другу красный, оливковый и оранжевый кластеры из предыдущей визуализации (рис. 3). Также в этот кластер ушли расположенные в непосредственной близости говоры д. Войница (121) и Коргуба (99);

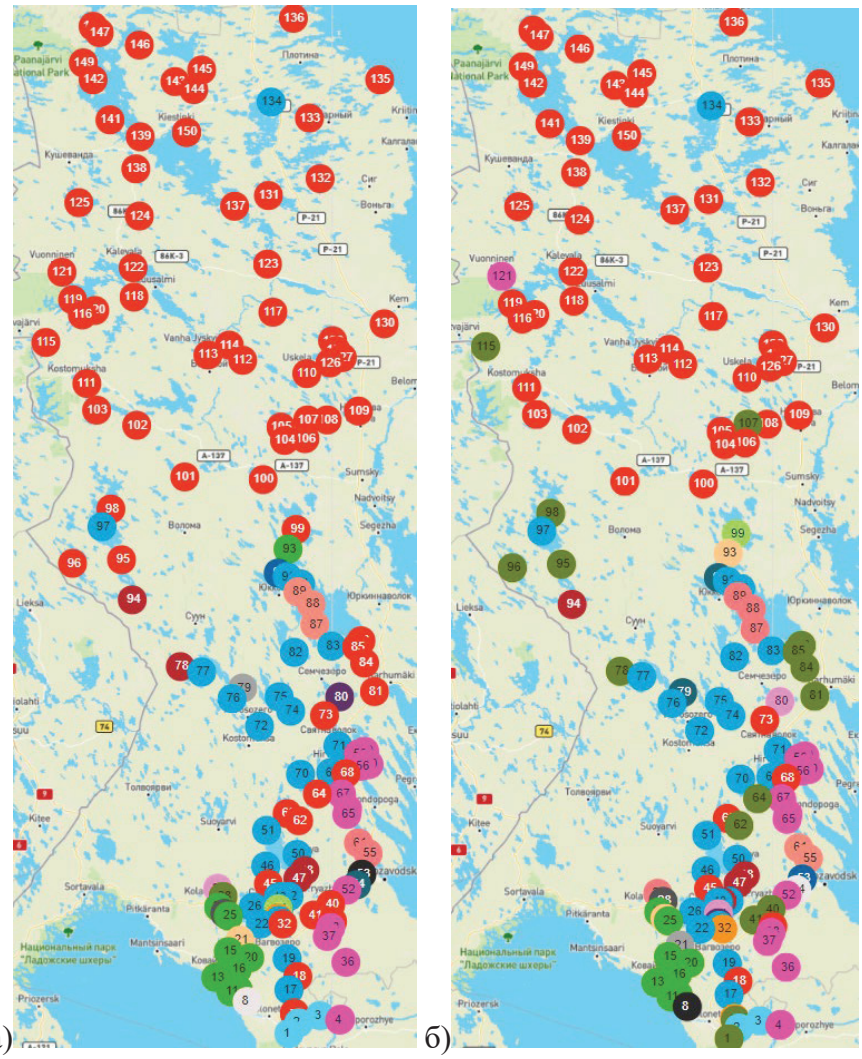
– южнее на территории южнокарельских и ливвиковских говоров сформировался небесно-голубой кластер (26 пунктов), практически полностью совпадающий с аналогичным кластером с предыдущей карты;

– на востоке от этого кластера оформился «людиковский» малиновый кластер (14 пунктов), также не обнаруживающий значительных отличий с соответствующим кластером с рис. 3;

– на юго-западе сохранился зелёный кластер (11 пунктов), однако, к нему присоединился говор д. Кузнаволок (93);

– коричневый кластер сократился до 4 пунктов.

Отдельные мелкие кластеры распались на самостоятельные или сохранились без каких-либо изменений.



а) б)

Рис. 4. Кластеризация а) центроидным методом и б) методом k-средних

В результате применения центроидного метода кластеризации сформировалось четыре больших и большое количество маленьких кластеров. В целом, в сравнении с более высокими уровнями кластеризации методом полной связи, отличия незначительны, но в данном случае не удалось проследить наличия каких-либо отличий внутри самых крупных кластеров, что, очевидно, связано с ограничением количества кластеров.

3. Кластеризация методом k-средних

Результат кластеризации методом k-средних зависит от выбора центроидов. Изначально пункты делятся на кластеры методом полной связи, затем в каждом полученном кластере вычисляются центроиды (объекты, сумма расстояний до

которых от остальных объектов в кластере минимальна), которые и принимаются в качестве начальных (эталонных). Далее высчитываются расстояния от всех «нецентроидных» элементов ко всем центроидам. Элементы распределяются в кластер к ближайшему центроиду (с наименьшим расстоянием). Во вновь образовавшихся кластерах центроиды перевычисляются заново. Если найдены новые центроиды, то процедура перераспределения элементов по кластерам повторяется. Такой метод даёт возможность одному говору на разных этапах анализа по-разному перераспределяться между кластерами.

В процессе сравнения результатов кластеризаций методами полной связи (рис. 3а) и k-средних (рис. 4б) были обнаружены следующие незначительные отличия:

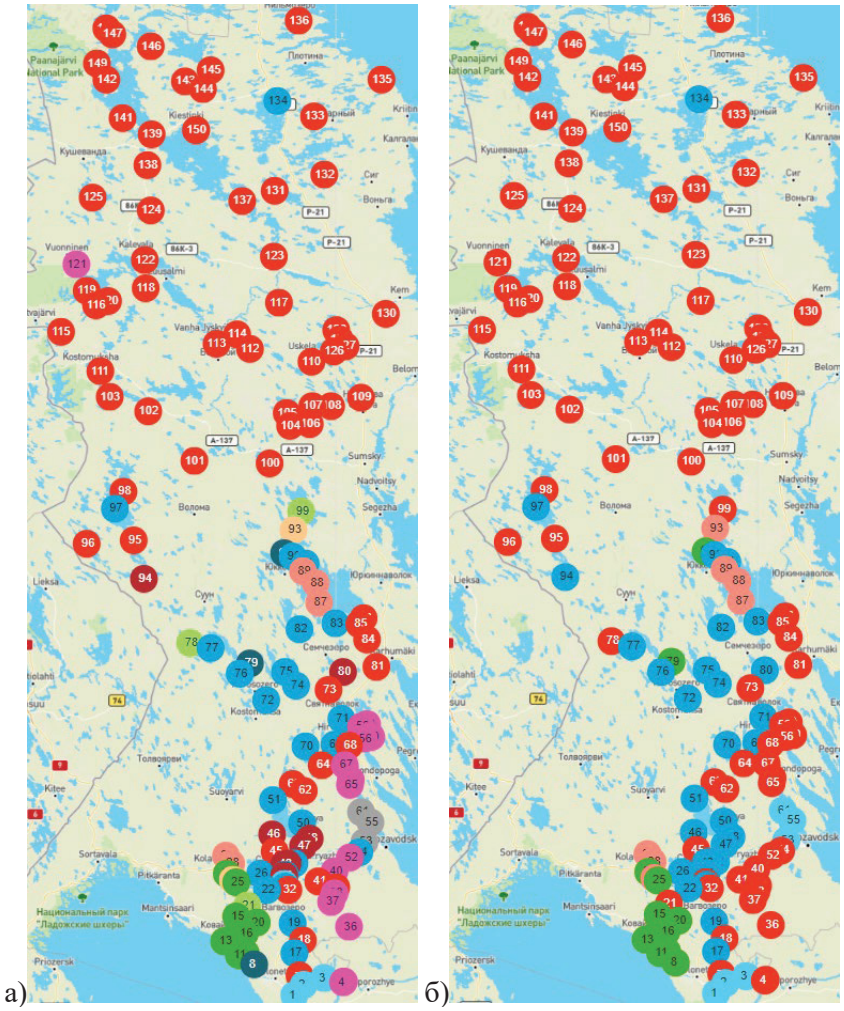
– в соседние более крупные кластеры кластеры перешли говоры д. Вешкелица (46), Ламбисельга (43), Соповаракка (112), Лубосалма (78);

– говор д. Сона (29), наоборот, отделился от соседнего кластера;

– говор д. Мегрега (1) из пурпурного кластера перешёл в более крупный оливковый. Ручной подсчёт отличий показал, что между говорами 1 и 24 в имеющемся материале содержится 5 отличий + 4 потенциальных отличия (отсутствие ответа), а между говорами 1 и 7 5 отличий или, соответственно, 9 совпадений + 4 потенциальных совпадения и 12 совпадений. Таким образом, переход можно считать правомерным.

Обсуждение и заключения

Как показало сравнение методов полной связи, центроидного и k-средних, результаты кластеризации не обнаруживают существенных отличий. Однако для более детального изучения междиалектных особенностей лучшим образом проявили себя методы полной связи и k-средних. Учёт положительных эффектов применения методов и параметров, а также анализ более высоких уровней кластеризации, позволяет получить итоговую карту распространения восходящих дифтонгов в говорах карельского языка Карелии (рис. 5).



а) б)

Рис. 5. Система восходящих дифтонгов в говорах карельского языка Карелии: кластеризации методом полной связи (а – 11 кластеров, минимальное расстояние между кластерами: 0,66; б – 5 кластеров, минимальное расстояние между кластерами: 0,76)

Распределение вариантов рефлексов **aa*, **ää*, *a-a*, *ä-ä* представлено на итоговой карте (рис. 5) следующим образом:

1) для большинства собственно карельских, а также отдельных ливвиковских сямозерских, видлицких, коткозерских, неккульских и рыпушкальских говоров для ударного слога характерно употребление дифтонгов *ua*, *iä* (красный кластер, рис. 5а). Некоторые отличия обнаруживает диалектное представительство восходящих дифтонгов, появившихся в языке вследствие стяжения: в партитиве единственного числа в именном словоизменении или в I инфинитиве в глагольном в севернокарельских диалектах обнаруживается дифтонг *yä* (*elyä*) в отличие от южнокарельского представительства (*iä – eliä*), а в южных мяндусельгских говорах – дифтонг *öä* (*elöä*);

2) людиковский кластер (малиновый, рис. 5а), в котором, как и в красном кластере, в ударном слоге используются дифтонги *ua*, *iä*, выделяется благодаря своей фонетической особенности, связанной с сохранением интервокального согласного в процессе альтернации согласных, что привело к заметно менее обильной дифтонгизации.

Близость кластеров подтверждается их слиянием на более высоких шагах кластеризации (красный кластер, рис. 5б). В него также вливаются говоры лимонного кластера (рис. 5а), находящиеся в так называемой зоне вибрации;

3) в поросозерских, паданских и южных мяндусельгских говорах употребляются дифтонги *oa*, *eä*, как и в коткозерских, сямозерских и большей части ведлозёрских ливвиковских говоров (небесно-голубой и коричневый кластер, рис. 5а, объединяющиеся в общий небесно-голубой кластер, рис. 5б);

4) для видлицких и тулмозерских говоров ливвиковского наречия (зеленый кластер, рис. 5а) характерно использование долгих гласных *aa*, *ää* / *ee*, очевидно, вторично развившихся из дифтонгов. Об этом свидетельствуют исследования начала прошлого века, в которых отмечается присутствие в тулмозерских и видлицких говорах дифтонгов *oa*, *eä* [17, 23]. При дальнейшей кластеризации (рис. 5б) в группу вливаются и демонстрирующие переходный характер говоры бирюзового кластера с рис. 5а;

5) схожую ситуацию: употребление долгих гласных *oo* / *aa*, *ee* обнаруживают ливвиковские северные тулмозерские, а также паданские и восточные ругозерские говоры собственно карельского наречия (розовый и песочный кластеры, рис. 5а), сливающиеся далее в общий кластер (розовый, рис. 5б);

6) отдельным среднелюдиковским и ливвиковским неккульским и рыпушкальским говорам (серый и голубой кластеры, рис. 5а) наряду с использованием дифтонгов *ua*, *iä* свойственно и явление их сужения до *uo*, *ie*. На более высоких уровнях кластеризации эти группы говоров объединяются (голубой кластер, рис. 5б).

Полученная методом кластеризации картина распределения восходящих дифтонгов в говорах карельского языка совпала с традиционным представлением, описанным в исследованиях по карельской фонетике и диалектологии (см. Введение), что, в свою очередь, указывает на правомерность дальнейшего применения описанной методики для решения недостаточно полно изученных проблем карельской диалектологии. Стоит особо отметить, что кластерные карты предоставляют довольно чёткую картину распространения диалектных вариантов анализируемого явления, дают возможность проследить их переходные зоны, на них максимально объективно представлены границы диалектных явлений, что планируется использовать в качестве основы в процессе разработки лингвистически обоснованной диалектной классификации карельского языка.

Список источников и литературы

1. Архангельский Т. А. Применение диалектометрического метода к классификации удмуртских диалектов // Урало-алтайские исследования. 2021. № 2. С. 7–20.

2. Баранцев А. П. Фонологические средства людиковской речи. Л.: Наука, 1975. 280 с.

3. Бубрих Д. В., Беляков А. А., Пунжина А. В. Диалектологический атлас карельского языка. Helsinki: SUS, 1997. 10 с.

4. Буреева Н. Н. Многомерный статистический анализ с использованием ППП “STATISTICA”. Нижний Новгород: [б. и.], 2007. 112 с. URL: http://www.unn.ru/pages/issues/aids/2007/57.pdf (дата обращения: 02.03.2022).

5. Разделы понятий. // Диалектная база карельского языка «Murreh». URL: http://murreh.krc.karelia.ru/sosd/concept_category (дата обращения: 27.01.2022).

6. Дугушкина Н. В. Обзор популярных методов кластеризации в машинном обучении // Наукосфера. 2020. № 7. С. 112–118.

7. Новак И. П. Базовая лексика карельского и вепсского языков в лингвогеографическом аспекте // Вестник угроведения. 2021. Т. 11. № 1. С. 90–101.

8. Основы финно-угорского языкознания: прибалтийско-финские, саамский и мордовские языки. М.: Наука, 1975. 347 с.

9. Сопоставительно-ономасиологический словарь диалектов карельского, вепсского и саамского языков / под ред. Ю. С. Елисеева, Н. Г. Зайцевой. Петрозаводск: КарНЦ РАН, 2007. 348 с.

10. Рягоев В. Д. Тихвинский говор карельского языка. Л.: Наука, 1977. 288 с.

11. Clustering Lexical Variation of Finnic Languages Based on Atlas Linguarum Fennicarum / Honkola T., Santaharju J., Syrjänen K., Pajusalu K. // Linguistica uralica. 2019. № 3. Pp. 161–184.

12. Dialectology for computational linguists / Nerbonne J., Heeringa W., Prokic J., Wieling M. // Similar languages, varieties, and dialects: A computational perspective. Cambridge: Cambridge University Press, 2021. Pp. 96–118.

13. Website “Dialektometrie Projekt” – Salzburg. URL: http://www.dialectometry.com/ (дата обращения: 10.01.2022).

14. Genetz A. Tutkimus Aunuksen kielestä. Helsinki: SKS, 1885. 194 p.

15. Genetz A. Tutkimus Venäjän Karjalan kielestä // Suomi, 1880. № 4. Pp. 1–248.

16. Website “Ison suomen kieliopin verkkoversio”. URL: https://kaino.kotus.fi/visk/etusivu.php (дата обращения: 20.01.2022).

17. Leskinen E. Tulemajärven murteen vokalismi. Helsinki: SKS, 1933. 138 p.

18. Leskinen H. Karjala ja karjalaiset kielentutkimuksen näkökulmasta // Karjala: historia, kansa, kulttuuri. Helsinki: SKS, 1998. Pp. 352–382.

19. Nerbonne J., Kretzschmar W. *Dialectometry++* // Literary and Linguistic Computing. 2013. Vol. 28. № 1. Pp. 2–12.

20. Novak I., Penttonen M. *Analysis of Karelian Dialect Division Based on Algorithmic Clustering* // Linguistica uralica. 2021. № 2. Pp. 81–101.

21. Rapola M. *Suomen kielen äännehistorian luennot*. Helsinki: SKS, 1966. 498 s.

22. *Website “Schweizerdeutsche Dialektometrie”*. URL: http://dialektkarten.ch/dmviewer/swg/ index.de.html (дата обращения: 15.01.2022).

23. Turunen A. *Lyydiläismurteiden äännehistoria. II*. Helsinki: SKS, 1950. 338 s.

24. Virtaranta P. *Über die donetzischen Konduši-Dialekt* // Finnisch-Ugrische Forschungen. 1973. № 40. Pp. 259–277.

25. Zaikov P. M. *Karjalan kielen murteet.* Петрозаводск: Изд-во ПетрГУ, 2017. 36 p.

References

1. Arkhangelskiy T. A. *Primenenie dialektometricheskogo metoda k klassifikatsii udmurtskikh dialektov* [Application of the dialectometric method to the classification of the Udmurt dialects]. *Uralo-altayskie issledovaniya* [Ural-Altaic Studies], 2021, no 2, pp. 7–20. (In Russian)

2. Barantsev A. P. *Fonologicheskie sredstva lyudikovskoy rechi* [Phonological means of Lyudic speech]. Leningrad: Nauka Publ., 1975. 280 p. (In Russian)

3. Bubrikh D. V., Belyakov A. A., Punzhina A. V. *Dialektologicheskiy atlas karel'skogo yazyka* [Dialectological Atlas of the Karelian language]. Helsinki: SUS Publ., 1997. 10 p. (In Russian)

4. Bureeva N. N. *Mnogomernyy statisticheskiy analiz s ispol'zovaniem PPP “STATISTICA”* [Multidimensional statistical analysis using the STATISTICA software package]. Nizhniy Novgorod: [w/p], 2007. 112 p. Available at: http://www.unn.ru/pages/issues/aids/2007/57.pdf (accessed March 02, 2022). (In Russian)

5. *Razdely ponyatiy* [Sections of notions]. *Dialektnaya baza karel'skogo yazyka «Murreh»* [Murreh Karelian Dialect Database]. Available at: http://murreh.krc.karelia.ru/sosd/concept_category (accessed January 27, 2022). (In Russian)

6. Dugushkina N. V. *Obzor populyarnykh metodov klasterizatsii v mashinnom obuchenii* [Overview of popular clustering methods in machine learning]. *Naukosfera* [Science-sphere], 2020, no. 7, pp. 112–118. (In Russian)

7. Novak I. P. *Bazovaya leksika karel'skogo i vepsskogo yazykov v lingvogeograficheskom aspekte* [Basic vocabulary of the Karelian and Vepsian languages in the linguistic and geographical aspects]. *Vestnik ugrovedeniya* [Bulletin of Ugric Studies], 2021, no. 11 (1), pp. 90–101. (In Russian)

8. *Osnovy finno-ugorskogo yazykoznaniya: pribaltiysko-finskie, saamskiy i mordovskie yazyki* [Fundamentals of Finno-Ugric linguistics: Baltic-Finnish, Sami and Mordovian languages]. Moscow: Nauka Publ., 1975. 347 p. (In Russian)

9. *Sopostavitel'no-onomasiologicheskiy slovar' dialektov karel'skogo, vepsskogo i saamskogo yazykov* [Comparative and onomasiological dictionary of dialects of the Karelian, Vepsian and Sami languages]. Ed by Yu. S. Eliseeva, N. G. Zaytzeva. Petrozavodsk: KarNC RAN Publ., 2007. 348 p. (In Russian)

10. Ryagoev V. D. *Tikhvinskiy govor karel'skogo yazyka* [The Tikhvin dialect of the Karelian language]. Leningrad: Nauka Publ., 1977. 288 p. (In Russian)

11. Honkola T., Santaharju J., Syrjänen K., Pajusalu K. Clustering Lexical Variation of Finnic Languages Based on Atlas Linguarum Fennicarum. *Linguistica uralica*, 2019, no. 3, pp. 161–184. (In English)

12. Nerbonne J., Heeringa W., Prokic J., Wieling M Dialectology for computational linguists. *Similar languages, varieties, and dialects: A computational perspective*. Cambridge: Cambridge University Press, 2021. pp. 96–118. (In English)

13. *Website «Dialektometrie Projekt».* Salzburg. Available at: http://www.dialectometry.com/ (accessed March 02, 2022). (In German)

14. Genetz A. *Tutkimus Aunuksen kielestä*. Helsinki: SKS, 1885. 194 p. (In Finnish)

15. Genetz A. Tutkimus Venäjän Karjalan kielestä. *Suomi*, 1880, no. 4, pp. 1–248. (In Finnish)

16. *Website "Ison suomen kieliopin verkkoversio"*. Available at: https://kaino.kotus.fi/visk/etusivu.php (accessed January 20, 2022). (In Finnish)

17. Leskinen E. *Tulemajärven murteen vokalismi*. Helsinki: SKS, 1933. 138 p. (In Finnish)

18. Leskinen H. Karjala ja karjalaiset kielentutkimuksen näkökulmasta*. Karjala: historia, kansa, kulttuuri*. Helsinki: SKS, 1998. pp. 352–382. (In Finnish)

19. Nerbonne J., Kretzschmar W. Dialectometry++. *Literary and Linguistic Computing*, 2013, no. 28 (1), pp. 2–12. (In English)

20. Novak I., Penttonen M. Analysis of Karelian Dialect Division Based on Algorithmic Clustering. *Linguistica uralica*, 2021, no. 2, pp. 81–101. (In English)

21. Rapola M. *Suomen kielen äännehistorian luennot*. Helsinki: SKS, 1966. 498 p. (In Finnish)

22. *Website "Schweizerdeutsche Dialektometrie"*. Available at: http://dialektkarten.ch/dmviewer/swg/index.de.html (accessed January 23, 2022). (In German)

23. Turunen A. *Lyydiläismurteiden äännehistoria. II*. Helsinki: SKS, 1950. 338 p. (In Finnish)

24. Virtaranta P. Über die donetzischen Konduši-Dialekt. *Finnisch-Ugrische Forschungen*, 1973, no. 40, pp. 259–277. (In Finnish)

25. Zaikov P. M. *Karjalan kielen murteet.* Petrozavodsk: Petrozavodsk State University Publ., 2017. 36 p. (In Finnish)

ИНФОРМАЦИЯ ОБ АВТОРАХ

Новак Ирина Петровна, старший научный сотрудник сектора языкознания, Институт языка, литературы и истории Карельского научного центра РАН (185910, Российская Федерация, Республика Карелия, г. Петрозаводск, ул. Пушкинская, д. 11), кандидат филологических наук.

novak@krc.karelia.ru

ORCID.ID: 0000-0002-9436-9460

Крижановская Наталья Борисовна, ведущий инженер-исследователь лаборатории информационных компьютерных технологий, Институт прикладных математических исследований Карельского научного центра РАН (185910, Российская Федерация, Республика Карелия, г. Петрозаводск, ул. Пушкинская, д. 11).

nataly@krc.karelia.ru

ORCID.ID: 0000-0002-9948-1910

ABOUT THE AUTHORS

Novak Irina Petrovna, Senior Researcher of the Department of Linguistics, Institute of Linguistics, Literature and History, Karelian Research Centre of the Russian Academy of Sciences (185910, Russian Federation, Republic of Karelia, Petrozavodsk, Pushkinskaya st., 11), Candidate of Philological Sciences.

novak@krc.karelia.ru

ORCID ID: 0000-0002-9436-9460

Krizhanovskaya Natalia Borisovna, Leading Research Engineer of the Laboratory for Information Computer Technologies, Institute of Applied Mathematical Research, Karelian Research Centre of the Russian Academy of Sciences (185910, Russian Federation, Republic of Karelia, Petrozavodsk, Pushkinskaya st., 11).

nataly@krc.karelia.ru

ORCID.ID: 0000-0002-9948-1910