УДК 81'286
DOI: 10.17223/19986645/64/3

Н.Г. Горлов, А.В. Кочановская, А.Н. Соболев

МЕТОДЫ КОМПЬЮТЕРНОЙ ЛИНГВОГЕОГРАФИИ В ИССЛЕДОВАНИИ ГРАНИЦ МЕЖДУ БЛИЗКОРОДСТВЕННЫМИ ЯЗЫКАМИ (НА ПРИМЕРЕ ДИАЛЕКТОВ ВОСТОЧНОЙ СЕРБИИ И ЗАПАДНОЙ БОЛГАРИИ)[1]

При помощи новейших методов компьютерной лингвистической географии решается теоретическая проблема кластеризации языковых и экстралингвистических данных и визуализируются объективные границы между близкородственными языками на лингвистических картах. Сгенерированные в ходе эксперимента пробные карты визуализируют кластеризацию рефлексов и резкость отграничения ареала их распространения от соседних, а также прямую корреляцию этих рефлексов с высотным расположением соответствующих населенных пунктов.

Ключевые слова: *южнославянские языки, сербские диалекты, болгарские диалекты, языковые границы, лингвистическая география, диалектометрия.*

Введение

Методы компьютерной лингвистической географии, разработанные к концу XX в. [1. S. 749–778], тогда же были применены к диалектам Восточной Сербии и Западной Болгарии [2], а результаты исследований вошли в синтетические труды по сербской и болгарской диалектологии (см. [3–5]) и общей ареальной лингвистике [6. S. 390–446]. Однако диалектометрических изысканий (в духе, например, [7, 8]) ни на материале данных приграничных южнославянских диалектов, ни на ином южнославянском материале в течение почти полутора десятилетий не предпринималось. Лишь в начале текущего столетия были осуществлены первые попытки верифицировать средствами математического анализа, в том числе и методами диалектометрии, традиционные диалектные классификации южнославянских языков – болгарского (см., например, [9]) и македонского [10]. Обзор двух последних и подобных им работ о языках Балканского полуострова, а также оценку их неоднозначных результатов см. в [11].

Цель предпринятого в настоящей статье исследования – впервые в славянской лингвистической географии применить новейшие компьютерные методы *кластеризации* и *визуализации* [12] к достоверным и количественно релевантным интралингвистическим и экстралингвистическим данным по

[1] Исследование выполнено при финансовой поддержке гранта РФФИ 18-512-76002 ЭРА_а «Изучение дивергенции и конвергенции традиций Центральных Балкан: реализация и перцепция».

смежным диалектам двух близкородственных южнославянских языков, сербского и болгарского. Настоящее изыскание носит пилотный экспериментальный характер и отражает помимо прочего ход ряда новых автоматических и полуавтоматических подготовительных и вспомогательных лингвистических работ, которые еще не являются рутинными для языковедов вообще и диалектологов в частности не только в России, но и в странах Западной Европы и США. Задачи исследования состоят в корректной машинной конвертации имеющихся в нашем распоряжении аналоговых первичных данных в цифровой формат, в разработке цифрового инструментария обработки и кластеризации первичных данных и в генерировании пробных лингвистических карт. Гипотезой теоретического исследования является предположение о том, что в результате применения методов систематизации, анализа, синтеза и визуализации географической дистрибуции *кластеров* языковых и экстралингвистических данных станет возможной картографическая *экспликация объективных границ* между близкородственными языками, в частности между сербским и болгарским. Практическим результатом исследования станет преодоление относительного неудобства и несовершенства печатных лингвистических атласов, в частности очевидных ограничений, накладываемых самим их форматом. Это и невозможность масштабирования карт и добавления в них новой информации (от новых пунктов до новых языковых данных), и трудоемкость сопоставления символов на лингвистических картах с данными из прилагающихся таблиц, и неудобства ручного наложения сетки пунктов или фоновой физико-географической карты на карты лингвистические, и невозможность создания большого количества комбинированных и диалектометрических карт, и т.д. В целом такой формат лишен динамичности, и работа с ним представляется излишне трудоемкой и крайне ограниченной в плане интерактивности. Эти и подобные им практические проблемы решены в результате разработки нового цифрового лингвогеографического инструментария.

Оцифровывание и дополнение первичных данных

Первичные данные для настоящего исследования, составляющие его электронную базу, были получены машинным оцифровыванием части материалов второго тома «Диалектологического атласа Восточной Сербии и Западной Болгарии (ДАВСЗБ)», содержащего вспомогательные и лингвистические карты [Sobolev 1998], а также дополнением этих материалов вручную новой релевантной экстралингвистической информацией. Работы проходили в два этапа, которые можно охарактеризовать как «общий» и «объектно-ориентированный». Каждый из двух этапов включает решение задач разных видов. Во-первых, это работы, необходимые для обеспечения материальной базы, не затрагивающие процесс дигитализации (например, нахождение координат пунктов, обработка, конвертация файлов, транслитерация текста и т.д.); во-вторых, это собственно оцифровывание материа-

лов ДАВСЗБ, OCR-распознавание (Optical Character Recognition) числовых и текстовых данных, в том числе таблиц; в-третьих, это регулярная контрольная сверка всех вносимых в базу сведений.

Первый этап представляет собой обработку вспомогательных данных, необходимых в дальнейшем исследовании в полном объеме, а основным его итогом стала электронная карта-сетка всех пунктов ДАВСЗБ. В ходе работы были решены следующие задачи:

1. Составлена таблица с названиями пунктов:

1.1) таблица пунктов ДАВСЗБ оцифрована в формат Microsoft Office Excel;

1.2) данные перепроверены, неточности при распознавании текста устранены.

2. Проведены работы по установлению географических координат пунктов ДАВСЗБ:

2.1) проанализировано современное состояние населенных пунктов, включенных в атлас;

2.2) определены географические координаты всех современных населенных пунктов, вошедших в сетку;

2.3) таблица названий ойконимов дополнена:

а) географическими координатами;

б) сведениями об изменениях в названиях поселений (если они были переименованы) со ссылками на источник;

в) сведениями об исчезновении или слиянии нескольких пунктов в один со ссылками на источник.

3. Первая электронная карта сетки пунктов:

3.1) сгенерирована;

3.2) перепроверена на наличие всех заявленных пунктов;

3.3) сопоставлена в приложении Photoshop с картой-сканом из аналогового издания;

3.4) выявлены, классифицированы и устранены все случаи несовпадения в расположении пунктов на аналоговой и электронной картах.

Для оцифровывания был использован постранично отсканированный вариант издания ДАВСЗБ в формате Portable Document Format (PDF), который при помощи программы Universal Document Converter был переформатирован в архив отдельных изображений по количеству страниц в документе. Это, во-первых, сняло ограничения при работе с форматом PDF – изображения были переведены в графический формат Joint Photographic Experts Group (JPEG), а во-вторых, облегчило работу с различными перечнями пунктов, вошедших в ДАВСЗБ и помещенных в нем на разные страницы издания. Для сбора списка населенных пунктов в один файл фрагменты таблицы, занимающей в аналоговом издании несколько страниц, были собраны в базовом приложении Microsoft Paint в единое изображение JPEG, которое при помощи online-конвертера Online PDF Converter (https://online2pdf.com/convert-jpg-to-excel) было переоформлено в таблицу Microsoft Excel, где каждой строке издания соответствовала одна

строка, а номер пункта, его название (а при наличии – и номера по «Болгарскому диалектологическому атласу») были разнесены по разным ячейкам. После конвертации в списке были обнаружены отдельные неточности, исправленные после перепроверки вручную: программой выборочно не распознавались графемы с диакритическими знаками (прописные и строчные š, č, ž), а также слова и фрагменты слов, сочетавшие знаки разных алфавитов (все ойконимы в ДАВСЗБ приведены в латинской транслитерации с добавлением знака ъ). Отформатированный и проверенный на отсутствие ошибок файл «Таблица населенных пунктов» стал базой для дальнейшей работы.

Для составления электронной географической карты таблицу Excel необходимо было дополнить географическими координатами пунктов, которые определялись нами главным образом посредством сервиса GPS Coordinates Google Maps Based (www.gps-coordinates.net) и в отдельных случаях сервисами Bing Maps Microsoft Based (www.bing.com/maps) и Latitude and Longitude Finder (www.latlong.net). Для поиска координат пунктов ячейки таблицы, включающие кириллические знаки, были транслитерированы в кириллицу полностью. В случае неудачи при поиске названий в вариантах, приведенных в аналоговом издании или транслитерированных в кириллицу, они транслитерировались в упрощенную латиницу для поиска «на английском языке». Параллельно с поиском координат исследовалось современное состояние пунктов: отдельные поселения не присутствуют на общедоступных картах, поэтому поиск сведений о них велся через официальные сайты общин и округов и открытые источники (например www.dimovo.bg, www.dimitrovgrad.rs/cir/onama). Часто поиск производился визуально по самой электронной карте, поскольку поиск по названию не давал результатов. Именно такой способ применялся при работе с ойконимами-омонимами: так как на исследуемой местности встречаются поселения с одинаковыми названиями (например Буковец, Главановци, Голеш, Градиште, Извор и др.), а в таблице они приведены в том порядке, в котором обследовались, т.е. территориально-последовательно, было принято решение искать пункт на карте в окружении тех, рядом с которыми он фигурирует в таблице.

В итоге в таблицу названий обследованных пунктов были добавлены данные по каждому из них – установлены координаты (широта и долгота помещены в отдельные ячейки), собраны сведения о слиянии пунктов в один (например, пп. № 305 и 306 объединены в с. Люлин) или их исчезновении с карты (например, п. № 530 исчез с наполнением Завойского озера).

По данным дополненной таблицы была сгенерирована первая электронная карта сетки пунктов, которая была перепроверена на наличие всех пунктов и отсутствие неточностей в названиях и номерах пунктов. Для оценки точности определения положения пунктов на карте было решено сопоставить карту издания ДАВСЗБ, существующую в формате JPEG, с электронной сеткой пунктов, нанесенной на план местности и преобразованной в формат Portable Network Graphics (PNG), как два изображения, для чего была выбрана про-

грамма Adobe Photoshop CC 2018. Для этого два файла открываются в программе как отдельные окна, посредством функций Слои > Создать дубликат слоя первое изображение дублируется внутри своего окна в два одинаковых слоя, один из которых перемещается в панель слоев в окне второго изображения. Таким образом, слой-дубликат карты-скана был перемещен к сгенерированной карте, разрешение и размер которой больше (8 МБ против 1,5 МБ), что снимало ряд технических трудностей с изменением размера общего файла в обратном случае. Файл «совмещенные карты» был преобразован в программе в смарт-объект, т.е. тип файлов в программах-иллюстраторах, допускающий редактирование растровых или графических изображений, помещенных в виде слоев. Была отрегулирована прозрачность слоев таким образом, что в смарт-объекте были четко видны знаки и номера пунктов. Задача состояла в том, чтобы изменить размер карты-скана, увеличив его до размеров сгенерированной карты так, чтобы они совпали по масштабу, чего нельзя было бы добиться, работая с бумажными носителями. При помощи функций Редактирование > Трансформирование > Масштабирование (Свободное трансформирование) меньший слой был перемещен и масштабирован до необходимых параметров.

В результате совмещения карт и анализа положений пунктов был установлен ряд несовпадений (для менее 10% от количества пунктов), которым была присвоена следующая кодировка:

– 0 – отсутствие пункта на сгенерированной карте (так помечались, например, второй пункт из двух, отсутствующий из-за слияния, или не найденный на карте пункт в месте полагаемой дислокации из-за значительного сдвига или в результате совпадения его координат с другим);

– 1 – сдвиг относительно оригинала (более ¾ несовпадений);

– 2, 4 – наложение координат (совпадение) разных пунктов – для двух разных случаев, когда по неустановленным причинам координаты пунктов совпали.

Все случаи несовпадений были пересмотрены – координаты перепроверены по альтернативным источникам, а изменения внесены в итоговую таблицу, ставшую основной при разработке цифрового инструментария.

1	Nr.	Name	Pesy	Latitude	Longitude	Nr. BDA	Комментарии		srtm3	srtm1	astergdem	gtopo30
362	361	Slavine		43.143825	22.846424	C4	Славиња, Србија	http://www.dimitrovgrad.rs/cir/o	731	731	740	758
363	362	Kamenica		43.135073	22.892794	C5			806	805	802	871
364	363	Krupъc	0	43.1308385	22.705330900000035	C6			950	952	951	903
365	364	Bolev Dol	1	43.118502	22.921622	C8			834	835	837	885
366	365	Brajk'ovci		43.122013	22.868142	C7	Брађевци, Србија	http://www.dimitrovgrad.rs/cir	756	754	756	768
367	366	Izatovci		43.1166867	22.88487310000005	C9			768	766	774	785
368	367	Gulenovci	4	43.121378	22.817856800000072	C10			1079	1079	1073	1151
369	368	Gorni Krivodol		43.1278923	22.96545530000003	C11			1258	1265	1261	1175
370	369	Visoki Odorovci		43.1025	22.816399999999993	C12	Височки Одоровци	http://www.dimitrovgrad.rs/cir	737	736	736	785
371	370	Vьlkovija	4	43.088299	22.91494	C13			852	851	854	876
372	371	Smilovci		43.0873128	22.845851100000004	C14			733	732	729	753
373	372	Moinci		43.0876836	22.88873060000003	C15			949	960	961	950
374	373	Petьrlaš		43.063610	22.787220	C16			797	807	813	794
375	374	Protopopinci		43.055801	22.861766499999993	C17			734	734	738	719
376	375	Mьzgoš	1	43.065310	22.903600	C18	Мазгош, Србија	http://www.dimitrovgrad.rs/cir	681	679	678	693

Рис. 1. Основная таблица-сетка обследованных пунктов (извлечение)

В ходе второго, «объектно-ориентированного», этапа было начато оцифровывание собственно лингвистических карт ДАВСЗБ. При этом решаются следующие задачи:

1. Изучение доступного программного обеспечения по электронному распознаванию текста с изображений и выбор оптимальной программы для дальнейшей работы.

2. Подготовка аналоговых материалов ДАВСЗБ для электронного распознавания.

3. Разработка единого алгоритма оцифровывания карты.

4. Оцифровывание перечней пунктов, входящих в состав каждой тематической карты, согласно единому алгоритму.

5. Составление электронных лингвистических карт.

Первоначально для оптического распознавания текста и преобразования данных в другой формат была выбрана программа Open OCR Cuneiform 2007, при помощи которой были оцифрованы карты ДАВСЗБ № 3, 7 и 20. В печатном атласе карты, посвященные рефлексам *dj, *o, *ь, используя наборы геометрических фигур, визуализируют ареальную дистрибуцию лингвистических признаков, обозначенных как form1, form2 и т.д., в обследованных населенных пунктах. В ходе работы с программой были выявлены ее существенные недостатки, в связи с чем был реализован переход к более современной ABBYY FineReader 14. Недостатки Cuneiform заключались, во-первых, в том, что программа не была способна к распознаванию таблиц, т.е. требовалось заранее обрезать изображение, перечень пунктов на котором дан в виде таблицы, в графическом редакторе по столбцам и лишь затем распознавать. Во-вторых, при распознавании текста программа порождала многочисленные ошибки типа 1 – !, 5 – 6, 2 – 3, 0 – 9, 7 – 1, 0 – О, 8 – В, пробел – нуль знака и т.д., что существенно затрудняло работу. В-третьих, в ней предполагается экспорт данных только в формат Microsoft Word. В отличие от Cuneiform, ABBYY распознает таблицы и изображения, имеет опцию экспорта в необходимый нам Microsoft Excel и опцию редактирования, т.е. сверки с оригиналом в самой программе, где результаты исправлений используются для машинного обучения. Разумеется, эта программа также допускает ошибки, однако в гораздо меньшем количестве и меньшего числа типов. Нами были замечены только: 0 – 9, 5 – 6, нуль знака – точка – запятая.

Подготовка материалов для данного этапа была осуществлена еще на первом этапе, когда посредством Universal Document Converter были получены страницы-изображения JPEG. Поскольку карты и перечни пунктов, нанесенных на них, даны стандартизованно, был разработан следующий алгоритм процесса оцифровывания:

1. В ABBYY:

1.1) «Открыть» > «Конвертация документов» > «Открыть в OCR-редакторе»;

1.2) провести сверку изображения и текстовых данных, представленных в разных частях окна программы, осуществить выверку опечаток и устранить их;

1.3) «Передать» документ, т.е. экспортировать его в формат Microsoft Excel.

2. В Microsoft Excel (выполнять отдельно для каждой формы (form 1 / form 2 и т.д.), которая представлена в материалах к аналоговой карте ДАВСЗБ каждая отдельным столбцом):

2.1) разделить данные одной строки на разные ячейки так, чтобы номера пунктов находились в отдельных ячейках:

2.1.1) выделить столбец, в котором, через запятые и / или пробелы или другие знаки табуляции записаны номера пунктов (как в издании ДАВСЗБ, где номера пунктов приведены подряд в один столбец по несколько в одной строке);

2.1.2) «Данные» > «Работа с данными» > «Текст по столбцам», указать формат «с разделителем», указать типы разделителей (точка, пробел, запятая) и отметить галочкой «считать последовательные разделители одним»;

2.1.3) ячейки всех столбцов перенести в один столбец и отсортировать по возрастанию.

2.2) отформатировать документ, вторично перепроверить данные;

2.3) сохранить с названием «карта № X».

Посредством разработанного единого алгоритма в будущем будет произведено оцифровывание всех лингвистических карт ДАВСЗБ.

Цифровой инструментарий и пробные карты

Цифровой лингвогеографический инструментарий и его функции разрабатывались на оцифрованном и дополненном материале ДАВСЗБ, представленном в виде таблиц, основная из которых, как изложено выше, содержит полный нумерованный список обследованных в исходном атласе населенных пунктов с координатами каждого из них, а также некоторую вторичную нелингвистическую информацию.

Разработка ведется на языке программирования R в среде разработки RStudio [12]. Основная особенность данного языка – его расширяемость с помощью свободно разрабатываемых и распространяемых библиотек, обеспечивающих работу специфических функций – так называемых пакетов. Одним из таких пакетов, сыгравшим решающую роль в выборе языка и среды для данного проекта, является Leaflet, который предоставляет мощный инструментарий для разработки интерактивных цифровых карт. Географической «основой» для них служат базовые карты, составленные участниками некоммерческого веб-картографического сообщества OpenStreetMap. Нами было принято решение использовать в качестве «основы» для нашей собственной разработки карту, созданную сообществом OpenStreetMap Sweden.

Первым этапом разработки стало написание скрипта на R, генерирующего карту и выводящего на нее данные из вышеупомянутой основной таблицы: каждый пункт был представлен на карте (в соответствии с его координатами) в виде круглого черного маркера, над которым был расположен его порядковый номер. Чуть позже скрипт был переработан таким образом, чтобы на карту выводились и другие данные из основной табли-

цы: второстепенная нелингвистическая информация о каждом пункте хранилась во всплывающих полях, появлявшихся при нажатии на маркер, а возле самих маркеров находились не только их порядковые номера, но и их названия.

На следующем этапе была поставлена цель создать инструмент поиска населенного пункта на сгенерированной карте, как по названию, так и по номеру. Возможности пакета Leaflet для решения этой задачи оказались недостаточными, в связи с чем было решено разработать отдельное приложение, которое объединяло бы в себе карту и поисковую систему. Для этого потребовался другой пакет языка R – Shiny, позволяющий создавать интерактивные веб-приложения. В результате было написано локальное веб-приложение, представлявшее из себя веб-страницу, на которой находилась сгенерированная интерактивная карта, а также, в виде двух отдельных панелей, поля поиска – по номерам населенных пунктов и по их названиям. При выборе того или иного пункта с помощью поисковика происходило автоматическое центрирование и приближение экстента карты к соответствующим координатам.

Необходимо отметить, что, начав работу над данным проектом, мы были заинтересованы во внедрении в инструментарий не только уже имевшихся у нас лингвогеографических данных из печатных атласов, но и иной, новой информации, в том числе внелингвистического характера. Первым шагом в этом направлении и следующим этапом стала разработка отображения на генерируемой карте данных о высоте каждого пункта над уровнем моря. Источником таких данных послужила информация, собранная международным исследовательским проектом по созданию цифровой модели высот «Радиолокационная топографическая миссия шаттла» (SRTM), а конкретнее, набор данных SRTM3, в котором общая площадь произведенной радарной топографической съемки делится на квадраты 90×90 м. Был написан отдельный R-скрипт, импортировавший в основную таблицу с перечнем населенных пунктов из ДАВСЗБ информацию о высоте каждого из них над уровнем моря. После этого для наглядного представления высотных данных на карте мы условно разделили их на шесть диапазонов:

– от 0 до 200 м над уровнем моря;
– от 200 до 400 м;
– от 400 до 600 м;
– от 600 до 800 м;
– от 800 до 1000 м;
– более 1000 м над уровнем моря.

Была создана шестичастная цветовая шкала, где каждый цвет соответствовал одному из выделенных диапазонов. Скрипт, генерирующий нашу карту, был переработан с учетом импорта перечисленной информации. В результате высотные данные были представлены на карте следующим образом: каждый круглый маркер, соответствующий населенному пункту из основной таблицы, был автоматически окрашен в один из цветов ше-

стичастной шкалы в соответствии с тем, в какой из выбранных шести диапазонов попадали указанные для этого пункта в таблице данные о его высоте над уровнем моря. Дополнительно эти непосредственные числовые данные были внедрены в вышеописанные всплывающие поля при каждом маркере. Кроме того, на «основу» карты был добавлен дополнительный слой теневой отмывки рельефа, созданный в рамках проекта OpenStreetMap. Следует также отметить, что на данной стадии (как и на всех последующих) при каждом маркере на карте сохранялся порядковый номер представляемого им населенного пункта, однако от выведения его названия возле маркера было решено отказаться в силу визуальной громоздкости такого представления.

Следующим этапом разработки лингвогеографического инструмента стали опыты по имплементации собственно лингвистических данных. Составленные ранее на основе трех карт ДАВСЗБ (карты № 3, 7 и 20) три таблицы формата XLSX, каждая из которых содержала информацию о том, какие формы представлены в исходной сетке и в каких населенных пунктах (в соответствии с нумерацией из нашей исходной таблицы) они встречаются, были сведены в три сетки встречаемости форм.

Для условного представления форм на созданной карте был выбран способ, аналогичный примененному в ДАВСЗБ: каждой форме (отражающей диалектный признак) соответствует условный знак, используемый в качестве маркера каждого пункта, в котором эта форма встречается. Для тех пунктов, где встречается более одной формы, был разработан особый маркер – черная точка, рядом с которой расположены два или более условных знака. Для пунктов, в которых не зафиксирована ни одна форма, также создан свой собственный маркер – черная окружность. Этот метод представления форм был внедрен в формирующий карту скрипт и отработан на таблице, содержащей данные из сетки карты № 3. Кроме того, разработанный метод представления форм был скомбинирован с описанным выше методом представления данных о высоте над уровнем моря. Таким образом, каждый условный знак при каждом населенном пункте на сгенерированной карте был автоматически окрашен в один из шести цветов в соответствии с имеющимися высотными данными об этом пункте. В настоящий момент ведется дополнительная работа над проверкой и уточнением сведений о высоте пп. над уровнем моря по традиционным источникам и над внесением необходимых коррективов.

Комбинационная карта, представленная на рис. 2, совмещает лингвистическую и внелингвистическую информацию, а именно сведения о рефлексах *dj (например в лексемах-рефлексах прасл. *medja ‘межа’) в обследованных пунктах (форма 1 – žd (mežda), форма 2 – dž (medža), форма 3 – ž (meža), форма 4 – g^{j} (meg^{j}a)), с одной стороны, и данные о высоте пунктов над уровнем моря – с другой.

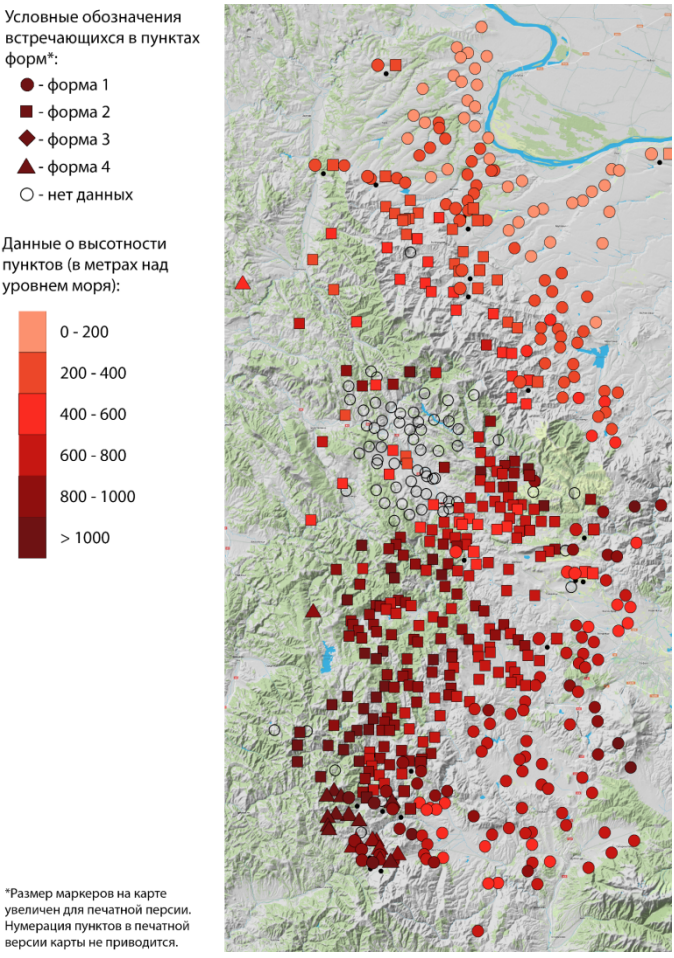

Рис. 2. Комбинационная карта с лингвистической и внешнелингвистической информацией

На момент написания статьи проводились также опыты по кластеризации и выводу на карту имеющейся у нас лингвистической информации в табличном формате различными комбинаторными методами. Так, был разработан метод градуированного представления межтабличной (и, соответственно, межсеточной) встречаемости форм в пунктах: каждый пункт на карте представлен круглым маркером, который окрашен в один из цветов из новой четырехчастной шкалы в соответствии с тем, в скольких из трех вышеупомянутых таблиц (соответствующих сеткам карт № 3, 7 и 20) содержится информация о том, что в этом пункте встречается какая-либо форма из трех, характерных для западноюжнославянского, в частности сербского, языкового ареала (т.е. **dj > dž, *ǫ > u, *ь > ə*): ни в одной, в одной из трех, в двух из трех, во всех трех таблицах.

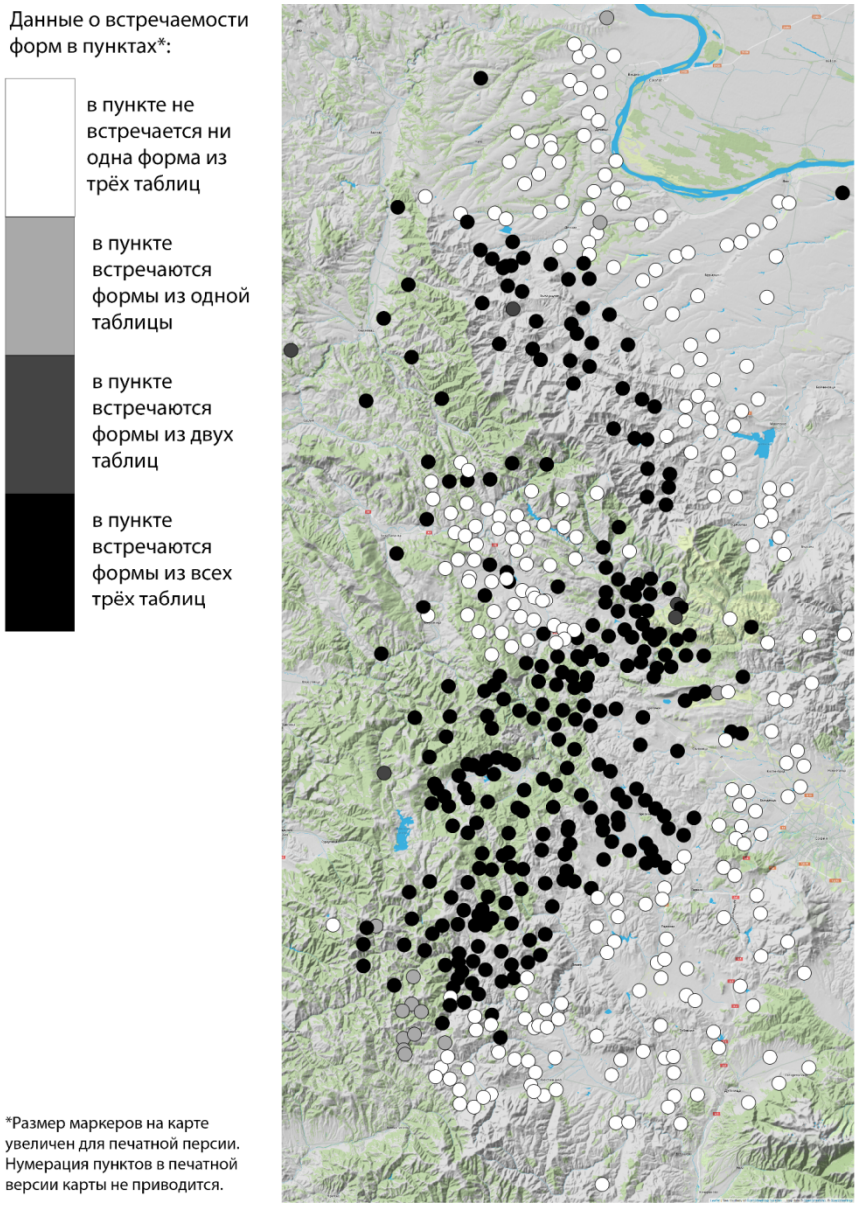

Рис. 3. Комбинационная карта с разноплановой лингвистической информацией

На этом материале сгенерирована комбинационная карта, представленная на рис. 3. В настоящий момент ведется дополнительная работа по автоматическому различению случаев полного отсутствия любых сведений из каких-либо пунктов от случаев отсутствия в этих пунктах именно искомых трех форм при возможном наличии других.

Заключение и перспективы дальнейших исследований

Полученные в ходе экспериментов результаты демонстрируют не только перспективность применения выбранных методов и возможность кластеризовать диалектные различительные признаки и внелингвистическую информацию, но и перспективы визуализации результатов на лингвогеографических картах. Сгенерированные экспериментальные карты демонстрируют кластеризацию рефлексов *dj > dž, *ǫ > u, *ь > ə и резкость отграничения ареала их распространения от соседних ареалов, а также их прямую корреляцию с высотным расположением соответствующих населенных пунктов в горном массиве Стара Планина по обе стороны государственной границы между Сербией и Болгарией. Рабочую гипотезу о возможности картографической *экспликации объективных границ* между близкородственными языками, в частности между сербским и болгарским, можно считать подтвержденной.

К ближайшим перспективам исследования относятся полное оцифровывание карт ДАВСЗБ, дополнение основной карты-сетки пунктов новой информацией, включая новые пункты, расширение возможностей интерактивного взаимодействия с картой и внутренними лингвогеографическими данными (на уровне как самой карты, выводимой в веб-приложение, так и отдельно встраиваемых в это приложение инструментов), увеличение объема и разнообразия этих данных, изучение и совершенствование способов их представления (так, встречаемость форм в пунктах может быть отображена не только с помощью условных знаков, но и посредством изоглосс, цветовой заливки определенных областей и комбинации этих методов), а также оптимизация хранения этих данных, их редактирования и оперативного обращения к ним. В дальнейшем к уже имеющимся и к новым количественно релевантным надежным данным можно будет применить самые современные методы статистического анализа [13, 14], что позволит надежно верифицировать лингвогеографические наблюдения над разграничением близкородственных языков.

Литература

1. *Putschke W., Neumann R.* Automatische Sprachkartographie // Dialektologie. 1. Halbband. Handbücher zur Sprach- und Kommunikationswissenschaft. Berlin : de Gruyter, 1982. S. 749–778.

2. *Sobolev A.N.* Sprachatlas Ostserbiens und Westbulgariens. Marburg, 1998. Bd. 2. 300 s.

3. *Павле И.* Целокупна дела. X/2. Расправе, студије, чланци. 2. О дијалектологији. Приредио Слободан Реметић. Сремски Карловци, Нови Сад : Издавачка књижарница Зорана Стојановића, 2018. 337 с.

4. *Български* диалектен атлас / отг. ред. И. Кочев. Обобщаващ т. 1–3: Фонетика. Акцентология. Лексика. София : Труд, 2001. 538 с.

5. *Български* диалектен атлас / отг. ред. М. Тетовска-Троева. Обобщаващ т. 4: Морфология. София : Проф. Марин Дринов, 2016. 247 с.

6. *Language* and space. Language mapping. An international handbook of linguistic variation / eds. by A. Lameli, R. Kehrein, S. Rabanus. Berlin ; New York : de Gruyter, 2010. Pt 1. XXII, 668 S.

7. *Goebl H.* Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie. Wien : Verlag der Österreichischen Akademie der Wissenschaften, 1982. 123 S.

8. *Goebl H.* Ansätze zu einer komputativen Dialektometrie // Dialektologie. 1. Halbband. Handbücher zur Sprach- und Kommunikationswissenschaft. Berlin : de Gruyter, 1982. S. 778–792.

9. *Prokić J.* Families and resemblances. PhD thesis. Groningen: s.n., 2010. (Groningen Dissertations in Linguistics 88). 196 p.

10. *Dombrowski A.* A Network Analysis of Macedonian Dialects (a Methodological Experiment). A paper presented at the 19th Biennial Conference on Balkan and South Slavic Linguistics, Literature and Folklore. April 25–27, 2014, University of Chicago, Illinois. 25 p.

11. *Русаков А.Ю., Морозова М.С.* Количественные исследования балканских языков и диалектов: достижения и перспективы // Съпоставително езикознание. 2020. 19 p. In print.

12. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

13. *Конер Д.В., Макарова А.Л., Соболев А.Н.* Статистический метод языкового профилирования носителя диалекта (на материале восточносербского идиома села Берчиновац) // Вестник Томского государственного университета. Филология. 2019. № 58. С. 17–33.

14. *Makarova A.L., Sonnenhauser B., Vuković T.* Corpus-based variation analysis in a Timok dialect. 34 p. manuscript.

Methods of Digital Linguistic Geography in Research on the Borders Between Closely Related Languages (Dialects of Eastern Serbia and Western Bulgaria)

Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology. 2020. 64. 42–55. DOI: 10.17223/19986645/64/3

Nikita G. Gorlov, Institute for Linguistic Studies of the Russian Academy of Sciences (Saint Petersburg, Russian Federation). E-mail: gorlov666@gmail.com

Anna V. Kochanovskaya, Saint Petersburg State University (Saint Petersburg, Russian Federation); University of Belgrade (Belgrade, Serbia). E-mail: kochanovskayaanna@yandex.ru

Andrey N. Sobolev, Institute for Linguistic Studies of the Russian Academy of Sciences (Saint Petersburg, Russian Federation); Philipps University of Marburg (Marburg, Germany). E-mail: sobolev@staff.uni-marburg.de

Keywords: South-Slavic languages, Serbian dialects, Bulgarian dialects, linguistic borders, linguistic geography, dialectometry.

In the study, the latest computer methods of clustering and visualization are applied to reliable and quantitatively relevant intralinguistic and extralinguistic data on adjacent cross-border dialects of two closely related South Slavic languages, Serbian and Bulgarian, for the first time in Slavic linguistic geography. The survey is a pilot experimental one. The objectives of the study are the correct machine conversion of the existing analog primary data into a digital format, the development of digital tools for processing and clustering the primary data, and the generation of trial combinatorial linguistic maps. It is assumed that as a result of applying the methods of systematization, analysis, synthesis, and visualization of the geographical distribution of clusters of linguistic and extralinguistic data, a cartographic explication of objective boundaries between closely related languages is possible. The primary data for the study were obtained by a machine digitizing of a part of the South Slavic dialect materials from the second volume of the *Dialectological Atlas of Eastern Serbia and Western Bulgaria* (DAESWB). The immediate tasks of digitalization included the compilation of an elec-

tronic geographic grid of the surveyed sites; the digitization of three linguistic maps from DAESWB (on reflexes of the Proto-Slavic *dj, *ǫ, *ь); the development of digital linguo-geographic tools and their functions in the R programming language in RStudio; the introduction to the study of extralinguistic information on the altitude of each site above sea level; the creation of combinatorial physical-geographical and linguistic maps using geometric shapes and color scales. A map was generated combining information about reflexes *dj with data on the altitude of sites above sea level, as well as a map clustering reflexes *dj>dž, *ǫ>u, *ь>ə. The generated experimental maps demonstrate the clustering of reflexes * dj>dž, *ǫ>u, *ь>ə and the contrast of the delimitation of their distribution area from neighboring areas, as well as their direct correlation with the altitude location of the corresponding settlements in the Stara Planina mountain range on both sides of the state border between Serbia and Bulgaria. The working hypothesis about the possibility of a cartographic explication of objective boundaries between closely related languages, in particular between Serbian and Bulgarian, can be considered confirmed. In the future, the study will reliably verify the linguo-geographical observations on the delineation of closely related languages.

References

1. Putschke, W. & Neumann, R. (1982) Automatische Sprachkartographie. In: Putschke W. et al. (eds) *Dialektologie. 1. Halbband. Handbücher zur Sprach- und Kommunikationswissenschaft*. Berlin: de Gruyter. pp. 749–778.

2. Sobolev, A.N. (1998) *Sprachatlas Ostserbiens und Westbulgariens*. Bd. II. Marburg: Biblion.

3. Ivić, P. (2018) *Celokupna dela* [Collected Writings]. X/2. 2. Sremski Karlovci; Novi Sad: Izdavačka knjižarnica Zorana Stojanovića.

4. Kochev, I. (ed.) (2001) *Balgarski dialekten atlas* [Bulgarian Dialect Atlas] General Volumes 1–3. Sofia: Trud.

5. Tetovska-Troeva, M. (ed.) (2016) *Balgarski dialekten atlas* [Bulgarian Dialect Atlas] General Volume 4. Sofia: Marin Drinov.

6. Lameli, A., Kehrein, R. & Rabanus, S. (eds) (2010) *Language and space. Language mapping. An international handbook of linguistic variation*. Pt. 1. Berlin; New York: de Gruyter.

7. Goebl, H. (1982) *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

8. Goebl, H. (1982) Ansätze zu einer komputativen Dialektometrie. In: In: Putschke W. et al. (eds) *Dialektologie. 1. Halbband. Handbücher zur Sprach- und Kommunikationswissenschaft*. Berlin: de Gruyter. pp. 778–792.

9. Prokić, J. (2010) *Families and resemblances*. PhD thesis. Groningen: [s.n.]. (Groningen Dissertations in Linguistics 88).

10. Dombrowski, A. (2014) *A Network Analysis of Macedonian Dialects (A Methodological Experiment)*. A paper presented at the 19th Biennial Conference on Balkan and South Slavic Linguistics, Literature and Folklore. 25–27 April 2014. University of Chicago, Illinois.

11. Rusakov, A.Yu. & Morozova, M.S. (2020) Kolichestvennye issledovaniya balkanskikh yazykov i dialektov: dostizheniya i perspektivy [Quantitative Research on Balkan Languages: Achievements and Prospects]. *S"postavitelno ezikoznanie*. In print.

12. R Foundation for Statistical Computing, Vienna, Austria. (n.d.) *R: A language and environment for statistical computing*. [Online] Available from: https://www.R-project.org/.

13. Koner, D.V., Makarova, A.L. & Sobolev, A.N. (2019) Linguistic/Dialectal Profiling of Dialect Speakers: The Method Presented on the Idiolect From BerčInovac, Eastern Serbia. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology*. 58. pp. 17–33. (In Russian). DOI: 10.17223/19986645/58/2

14. Makarova, A.L., Sonnenhauser, B. & Vuković, T. (n.d.) *Corpus-based variation analysis in a Timok dialect*. A manuscript.