Slovenski jezik / Slovene Linguistic Studies 17 (2025) | https://doi.org/10.3986/17.1.07 | COBISS: 1.01 | CC BY SA 4.0
137–160

Анастасия Сергеевна Кузьмина [Anastasija Sergejevna Kuzmina]
Высшая школа экономики, Россия
askuzmina5487@gmail.com | https://orcid.org/0009-0009-5647-5105

Арсений Владимирович Манусов [Arsenij Vladimirovič Manusov]
Высшая школа экономики, Россия
arseniymanusov@gmail.com | https://orcid.org/0009-0007-1152-1353

Михаил Николаевич Саенко [Mihail Nikolajevič Sajenko]
Институт славяноведения РАН, Россия
michail.sajenko@yandex.ru | https://orcid.org/0000-0002-5829-7527

Классификация славянских языков на материале Общеславянского лингвистического атласа: пробный шар

В статье предлагается инновационный подход к классификации славянских языков, базирующийся на данных, представленных в Общеславянском лингвистическом атласе. В основе подхода лежит попарное сравнение рефлексов праславянских гласных из 780 населённых пунктов. К матрице близостей говоров применяется алгоритм иерархической кластеризации, позволяющий разделить славянский континуум на 15 ареалов. К каждому ареалу предлагается не только формальное описание, но и теоретическое осмысление.
КЛЮЧЕВЫЕ СЛОВА: славянские языки, классификация, диалектометрия, Общеславянский лингвистический атлас

Članek predlaga inovativen pristop h klasifikaciji slovanskih jezikov na podlagi podatkov, predstavljenih v Slovanskem lingvističnem atlasu. Pristop temelji na parni primerjavi refleksov praslovanskih samoglasnikov iz 780 lokacij. Za ugotavljanje narečne bližine je uporabljen hierarhični algoritem združevanja, ki omogoča razdelitev slovanskega območja na 15 območij. Pri vsakem območju se ob formalnem navaja tudi teoretičen opis.
KLJUČNE BESEDE: slovanski jeziki, klasifikacija, dialektometrija, Slovanski lingvistični atlas

This article proposes an innovative classification methodology for Slavic languages based on data from the The Slavic Linguistic Atlas (OLA). The approach employs pairwise comparison of Proto-Slavic vowel reflexes across 780 settlements. Applying hierarchical clustering algorithms to a matrix of dialect proximity measures, we have identified 15 distinct linguistic areas within the Slavic continuum. For each delineated area, we provide both formal linguistic descriptions and theoretical interpretations.
KEYWORDS: Slavic languages, classification, dialectometry, The Slavic Linguistic Atlas

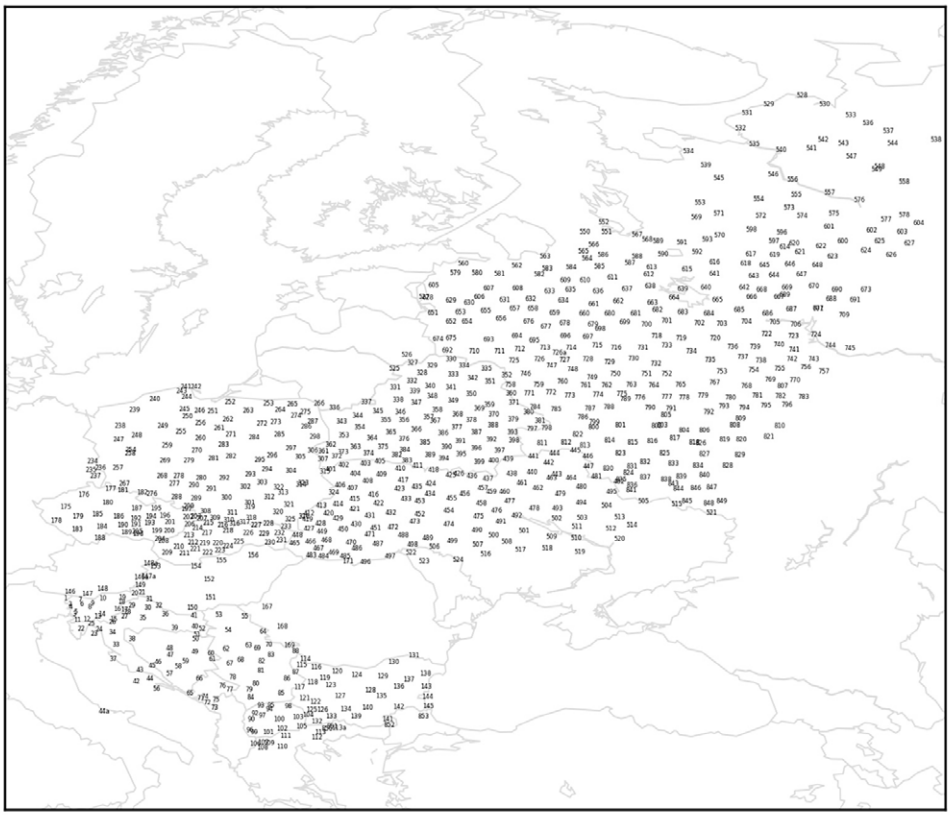
1 Вводные замечания

Первые попытки создать научную классификацию славянских языков принадлежат перу «отца славистики» Й. Добровского. Последующие двести лет развития науки породили десятки разнообразных классификаций, строящихся на различных основаниях, включающих в себя разное количество языков и разделяющих эти языки на разное количество групп и подгрупп [Saenko 2022].

К сожалению, большинство этих классификаций пыталось работать с данными только (или, по крайней мере, преимущественно) литературных языков, не учитывая говоров и того факта, что славянские языки представляют собой несколько континуумов, в рамках которых границы диалектов не всегда совпадают с границами литературных языков.[1] Кроме того, статус литературного языка во многом является результатом исторической случайности: если бы история пошла по другому пути, вполне возможно, что славянский континуум (с теми же изоглоссами) сейчас официально делился бы на большее или наоборот меньшее количество литературных языков, что напрямую сказалось бы на классификации. В то же время, диалекты, обычно считающиеся частью одного языка, иногда могут демонстрировать старые фонетические черты, характерные для другого языка. Таким образом строго лингвистическая классификация славянских языков должна строиться в первую очередь на данных говоров без учёта того, к каким языкам эти говоры относят по экстралингвистическим критериям.

Подходящий материал для такой классификации был собран в рамках проекта «Общеславянский лингвистический атлас» (далее ОЛА). Работа над ОЛА началась в 1958 г. после IV съезда славистов в Москве. Сбор полевого материала вёлся преимущественно в 1960–70-е гг. Сеть атласа включает в себя 830 населённых

[1] Похвальным исключением являются работы М. Шекли, например [Šekli 2018].



КАРТА 1: 780 пунктов ОЛА, используемых в работе (об исключении 50 пунктов из-за недостатка данных см. ниже)

пунктов: 1–21, 146–149 – словенские[2] (всего 25); 22–88, 146а–148а, 150–153, 168, 169 – сербохорватские (всего 77); 90–113а – македонские (всего 24); 114–145, 167, 850-853 – болгарские (всего 37); 175–206, 299 – чешские (всего 33); 154–156, 208–232 – словацкие (всего 28); 234, 236 – нижнелужицкие; 235, 237 – верхнелужицкие; 207, 238–240, 246–298, 300–326 – польские (всего 84); 241–245 – кашубские (всего 5). Восточнославянские пункты распределяются по странам следующим образом: 327–400 – в Белоруссии (всего 74); 171 в Румынии, 233 в Словакии, 401–524 на Украине (всего 124); 525–558, 560, 562–593, 596–598, 600–648, 651–671, 673–707, 709–716, 718–765, 767–841, 843–849 в России (всего 313); при этом языковые границы не совпадают с государственными (подробнее см. ниже в разделе 4).

[2] Языковая маркировка отдельных пунктов наша. В самом ОЛА она отсутствует.

Выпуски атласа выходят в двух сериях, фонетико-грамматической и лексико-словообразовательной. В первой из них к настоящему времени напечатано 9 томов: “1. Рефлексы *ě” (1988 г.), “2а. Рефлексы *ę” (1990 г.), “2б. Рефлексы *ǫ” (1990 г.), “3. Рефлексы *ьr, *ъr, *ьl, *ъl” (1994 г.), “4а. Рефлексы *ъ, *ь” (2006 г.), “4б. Рефлексы *ъ, *ь. Вторичные гласные” (2003 г.), “5. Рефлексы *o” (2008 г.), “6. Рефлексы *e” (2011 г.), “9. Рефлексы *tort, *tolt, *tert, *telt, *ort, *olt” (2019 г.). Отдельно вышел том с болгарскими материалами к выпускам 1–4 (2015 г.)[3]. Тома, посвящённые истории прочих гласных и согласных, пока ещё не выпущены и информацию на эту тему можно почерпнуть из данных ОЛА лишь в ограниченном объёме.

Несмотря на свои внушительные размеры, сетка пунктов ОЛА далека от идеала. Она обладает недостаточной густотой в некоторых ареалах с сильным диалектным дроблением (например, словенском[4]) а также в ареалах со смешанным населением. Кроме того, следует помнить, что материалы ОЛА отражают состояние на 1960–70-е гг. и не позволяют отследить даже поздних случаев смещения языковых границ, например, обмен населением между ПНР и СССР в 1944–1946 гг.[5], не говоря уже о более ранних миграциях, таких как русская и украинская колонизация Дикого поля[6] или переселения болгар, сербов и хорватов после захвата турками Балкан. Таким образом, демонстрируемые картами ОЛА изоглоссы далеко не всегда восходят к праславянскому состоянию. Тем не менее, обследование по программе ОЛА – это единственный массовый сбор материала говоров всех славянских языков по унифицированной программе, что делает данные ОЛА по-настоящему незаменимыми и заслуживающими всестороннего изучения.

---

[3] Болгарская комиссия ОЛА временно выходила из проекта в связи со своей радикальной позицией по македонскому вопросу.

[4] С.Б. Бернштейн изначально предлагал, чтобы в атлас вошло 40 словенских пунктов. Т. Логар считал, что это число нужно увеличить до 50, но в конечном итоге по различным причинам это количество было сокращено до 25 [Kenda-Jež 2012: 62–63].

[5] Так, в польских пунктах 238–240, 252, 253, 257, 267, 268 и 276 живут выходцы с восточных кресов.

[6] Например, в Воронежской области как русские, так и украинские пункты ОЛА возникли в XVII–XVIII вв.: 828 (Хреновое) в 1680-е гг.; 834 (Новая Ольшанка) в 1665 г.; 840 (Истобное) между 1645 и 1650; 847 (Новая Осиновка) в конце XVII в.; 849 (Шапошниковка) в 1750-е гг. [Прохоров 1973: 339, 199, 112, 200, 344].

2 Славянские говоры и диалектометрия

Диалектометрический (количественный) подход к классификации отдельных славянских языков и говоров начал применяться с конца XX в. Классификация, разработанная Н.Н. Пшеничновой [1996], предполагала разделение русских говоров на единицы четырех уровней: первые два уровня, объединяющие большое количество населённых пунктов, отображали русский ареал дискретно, тогда как третий и четвертый уровни благодаря высокой детализации позволяли представлять ареал континуально. При описании каждого населённого пункта были взяты все лингвистические материалы, используемые для составления карт ДАРЯ за исключением статистических карт. Также на основе материалов ДАРЯ была предложена классификация русских диалектов [Марченко, Ронько 2025]. В работе исходная матрица признаков была превращена в матрицу близостей населённых пунктов и затем уменьшена с помощью метода многомерного шкалирования. Полученные значения были визуализированы на карте, что позволило выделить 6 групп говоров и определить диалектные признаки, делящие русские говоры на запад и восток.

В работе [Кузьмина, Манусов 2024] с помощью диалектометрических методов был описан весь восточнославянский ареал. В основе классификации лежали карты из сборника «Восточнославянские изоглоссы», которые были перерисованы в электронном формате и трансформированы в бинарную матрицу, где каждому пикселю карты соответствовало наличие или отсутствие конкретной диалектной черты. Далее к полученной матрице был применён алгоритм *k*-средних, в результате чего были созданы две кластеризации: на 5 (макроуровневая) и на 25 (микроуровневая) кластеров. Для каждого кластера был предложен список дифференциальных диалектных признаков, вносящих наибольший вклад в их формирование. Помимо дискретного кластерного метода был также применен континуальный, построенный с применением Евклидова расстояния.

Используя материалы проекта «Идеографски диалектен речник на българския език», группа исследователей попыталась получить классификацию болгарских говоров при помощи расстояния Левенштейна. В качестве материала использовалось 39 языковых черт в 156 словах, записанных в 197 населённых пунктах каждое [Houtzagers et al. 2010]. Такая методика имеет свои ограничения: во всех изучаемых пунктах должен быть представлен одинаковый набор лексем, что при работе с материалом не отдельных языков, а таксонов становится проблематичным. Кроме того, этот метод критиковался за то, что любая разница между сопоставляемыми единицами оценивается одинаково. В качестве альтернативы П. Хеггарти предложил метод, при котором диалектные формы сравниваются

не напрямую, а с общей праформой и диалектные рефлексы получают разный вес в зависимости от своего качества [McMahon 2005: 210–239].

Количественным методом является также лексикостатистика, однако большинство работ по славянской лексикостатистике ограничивается данными литературных языков.[7] Важным шагом к изучению в этом ключе также говоров стала работа А.В. Дыбо и А.С. Касьяна, где наряду со списками Сводеша литературных славянских языков были использованы данные ряда говоров [Kushniarevich et al. 2015]. Полностью на диалектных данных (были взяты 25 славянских говоров) основана работа М.Е. Васильева и М.Н. Саенко [Васильев, Саенко 2020]. На материалы корпусов и списков Сводеша опирается также И. Афанасьев. Особого внимания заслуживает его работа [Afanasev 2023]. В ней автор провёл сравнительный анализ двух принципиально различных подходов к определению расстояний между языками: систем типа «черный ящик», показывающих языковую близость без возможности интерпретации, и «языково-независимых» систем, обеспечивающих прозрачную обработку данных, но игнорирующих специфику отдельных языков. Помимо сравнения существующих подходов он также предложил новую «антипрозрачную» систему, учитывающую преимущества обоих методов, и применил ее к трём восточнославянским лектам: говорам с. Хиславичи, Мегра и Белогорное.

Следует подчеркнуть, что классификации, полученные с помощью диалектометрических методов в нашей работе, не предлагаются в качестве замены существующих классификаций, основанных на качественных подходах. Диалектометрия в исследовании – лишь один из способов увидеть связи и близости между лектами, которые могут остаться незамеченными без статистического анализа.

3 Методология и данные

3.1. Материал

Используемый материал представлен 6 томами фонетико-грамматической серии ОЛА (1, 2а, 2б, 4а, 5 и 6), посвящёнными континуантам следующих праславянских гласных: *ě, *ę, *ǫ, *ъ и *ь, *о, *е. Во всех томах приведены реконструированные

[7] Если говорить о лексикостатистике не в «сводешевском» понимании, то установление близости языков на основании сравнения параллельных текстов, предложенное независимо В. Маньчаком [Mańczak 2006] и С.А. Старостиным [Старостин 2007], в целом трудно применить к диалектному материалу.

лексемы праязыка, содержащие рассматриваемые фонемы. Для каждой лексемы приводится список длиной 830 (количество обследованных населённых пунктов) форм, наследующих исходной лексеме. В населённом пункте может быть засвидетельствовано как несколько, так и ни одной удовлетворяющей формы. Одна форма может фиксироваться в нескольких подряд идущих пунктах. В качестве примера приводятся данные по населённым пунктам 715–719 карты 13 (*č/ь̯/stь̯) тома на редуцированные гласные (пункт 717 в атласе отсутствует, в нумерации бывают пробелы):

715–716	čes’t’
718	č’es’, č’is’
719	

Тома были преобразованы в таблицы, из форм нами были извлечены непосредственно сами рефлексы – звуки и звукосочетания, развившиеся из фонемы праязыка. Карты из тома 4а в дальнейшем были разделены на две таблицы, в зависимости от того, рефлексы какого из редуцированных гласных там представлены. Так, после разметки рассматриваемый фрагмент таблицы тома на *ь, соответствующий приведённому выше примеру, приобрёл следующую структуру:

	(13) č/ь̯/stь̯
715	e
716	e
718	e / i
719	нет

Населённые пункты, для которых хотя бы в одном из томов больше половины значений отсутствовало, были исключены из рассмотрения, в результате чего общее количество пунктов сократилось до 780. Были исключены следующие населённые пункты: 23, 34, 44a, 64, 93, 106–109, 112, 113, 139, 140, 142, 145, 167, 207, 253, 307, 327, 534, 538, 539, 550, 553, 558, 566, 598, 601, 615, 629, 630, 645, 669, 676, 677, 731, 735–737, 739–741, 755, 767, 768, 802, 808, 852, 853.

Важно отметить, что одни рефлексы встречались в двух и более томах, а другие — оказывались специфичными только для одного тома. Например, рефлекс *wǫ* был зафиксирован лишь трижды в томе на *o, в то время как рефлексы *a* или *i* были представлены в значительном количестве во всех томах. Вместе с тем все рефлексы демонстрировали структурную схожесть, что послужило основанием для объединения рефлексов разных томов в один набор данных.

Однако принятая методология имеет ряд ограничений: во-первых, теряется информация о связи с исходной фонемой, а, во-вторых, при включении нового материала, который будет структурно отличаться от имеющегося, например, из тома атласа на сочетания вида *TorT, методология будет нуждаться в совершенствовании.

3.2 Рефлексы

При рассмотрении всех встретившихся рефлексов (всего 344) нами была предпринята попытка осмыслить их структуру, чтобы далее было возможно производить сравнение. Формальная схема всех рефлексов была представлена следующим образом:

1. Группа гласного, имеющая следующую внутреннюю структуру, либо её отсутствие:

1.1. Один гласный из таблицы (Таблица 1). Каждый гласный охарактеризован по подъему (от самого верхнего, 1-го, до самого нижнего, 7-го), ряду (от самого переднего, 1-го, до самого заднего, 11-го), степени реализации (0 при нулевой реализации, 1 – для неслоговых частей дифтонгов, 2 – для гласных полного образования), огубленности (1 при наличии и 0 при отсутствии) и назализации (2 при наличии и 0 при отсутствии).

1.2. Подстрочная дужка на два символа либо её отсутствие.

1.3. Долгота (:) или полудолгота (·) либо их отсутствие.

2. i̯ в интервокальной позиции (в рефлексе *i̯i̯e*) либо его отсутствие.

3. Группа гласного, имеющая внутреннюю структуру, аналогичную пункту 1. списка, за исключением подпункта 1.2. либо её отсутствие.

4. Один носовой согласный (m, n, ŋ, ń, ŋ́, ^{m}, ^{n}) либо его отсутствие.

	1	2	3	4	5	6	7	8	9	10	11
1	i			и		ы			ъ		u
2		ɪ	у		ɐ̞						
3		ɐ̝	ɘ							ɵ	
4			e	ø		ɘ			o		
5				ɛ		ʌ		ɔ			
6					a̟		ɑ̟				
7						a					

РИСУНОК 1: Треугольник гласных в транскрипции ОЛА

Звук	Реализация	Подъём	Ряд	Огубленность	Назализация
0[8]	0	4	6	0	0
a	2	7	6	0	0
a̯	2	6	5	0	0
a̯̯	2	6	7	1	0
ɛ	2	5	4	0	0
ʌ	2	5	6	0	0
ɔ	2	5	8	1	0
e	2	4	3	0	0
e̯	2	4	6	0	0
e̯̯	1	4	6	0	0
ø	2	4	4	1	0
ø̯	2	3	2	1	0
ə	2	4	6	0	0
ə̯	2	4	6	0	2
o	2	4	9	1	0
o̯	2	4	6	1	0
ɐ̯	2	3	2	0	0
ɘ	2	3	3	0	0
ɐ̯̯	2	3	5	0	0
o̯̯	2	3	10	1	0
ɪ	2	2	2	0	0
y	2	2	3	1	0
i	2	1	1	0	0
i̯	1	1	1	0	0
j̑[9]	1	1	1	0	0
ľ	1	1	1	0	0
ń	1	1	1	0	0
i̯̯	1	1	1	0	2
i̯̯̯	2	2	6	0	0
ʉ	2	1	4	1	0
ы	2	1	6	0	0
ъы	2	1	9	0	0
u	2	1	11	1	0
u̯	2	2	6	1	0
u̯̯	1	2	6	1	0
v	1	2	6	1	0
w	1	2	6	1	0
ą	2	7	6	0	2
ą̯	2	6	5	0	2
ą̯̯	2	6	7	1	2
ę	2	5	4	0	2
ę̯	2	4	3	0	2
ę̯̯	2	3	2	0	2
į	2	1	1	0	2
ǫ	2	4	9	1	2
ǫ̯	2	3	10	1	2
ʊ̯	2	2	3	1	2
ų̯	2	1	11	1	2
ʊ	2	2	10	1	0

ТАБЛИЦА 1: Классификация гласных рефлексов

[8] Было принято решение принять 0 за центральный гласный (ə) с нулевой реализацией.

[9] В силу сложности корректного определения веса этих элементов неслоговые части дифтонгов были разделены на два типа: *i̯ = j = j̯* и *u̯ = w = v*. Кроме того, звуки *ľ* и *ń*, иногда развивающиеся на месте *j*, получили тот же вес, что и сам *j*.

3.3. Попарное сравнение

На следующем этапе исследования было проведено попарное сравнение имеющихся рефлексов, направленное на установление условных расстояний между ними. Сравнение двух рефлексов заключалось в соотнесении характеристик составных частей рефлексов. Характеристики гласных приведены в таблице 1; долгота, полудолгота и отсутствие долготы получили вес 2, 1 и 0 соответственно; интервокальный *i̯* при наличии давал вес 1, при отсутствии – 0; подстрочная дужка не учитывалась. В ходе такого сравнения рефлексы были представлены как векторы равной длины, чтобы между ними можно было вычислить Евклидово расстояние (схема 1), которое и стало конечной мерой дистанции между парой рассматриваемых рефлексов. Данный подход позволил перейти от качественного описания рефлексов к их количественной характеристике, что повысило объективность сравнительного анализа.

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

СХЕМА 1: Формула подсчёта Евклидова расстояния

Для сопоставления структурно различных единиц, а именно монофтонгов и дифтонгов, например, *a* и *a:i*, был разработан специальный алгоритм, в основе которого лежат два сравнения. Первое – прямое, при котором первый гласный дифтонга сравнивается с единственным гласным монофтонга, а второму гласному дифтонга в пару ставится вектор $\overline{q} = \overline{v} - \overline{c}$, где $\overline{v}$ – это вектор характеристик второго гласного, а $\overline{c}$ – это вектор такой же длины, что и $\overline{v}$, все значения которого равны некоторой константе *C*. Второе сравнение — обратное, при нём единственный гласный монофтонга будет сравниваться со вторым гласным дифтонга. Выбирается то сравнение, которое даст меньший результат. Таким образом, в приведённом примере меньший результат обеспечит прямое сравнение: соотнесение первого элемента дифтонга *a:* с монофтонгом *a*, а *i* — с $\overline{q}$; итоговое значение близости в данной паре составит 4,8990. Значения в матрице близостей (Таблица 2) лежат в диапазоне от 0 при сравнении одинаковых рефлексов до 17,2047 между рефлексом *uu* и рефлексами *in*, *iń*, *iŋ*, *i̯n*, *i̯ń*, *i̯ŋ*, *i̯ⁿ*. Такой подход позволяет верно устанавливать соответствие элементов одного рефлекса элементам другого, «штрафуя» за неправильные сравнения.

Два дифтонга между собой сравниваются аналогично: из прямого и обратного сравнений выбирается такое, которое обеспечит меньший результат.

	0	**a**	**a:**	**a:i**	**a:i̯**	**a:u̯**
0	0	3	3,6055	5,7446	5,7446	5,7446
a	3	0	2	4,8990	4,8990	4,8990
a:	3,6055	2	0	4,4721	4,4721	4,4721
a:i	5,7446	4,8990	4,4721	0	1	5,2915
a:i̯	5,7446	4,8990	4,4721	1	0	5,1966
a:u̯	5,7446	4,8990	4,4721	5,2915	5,1966	0

ТАБЛИЦА 2: Фрагмент квадратной матрицы попарных сравнений

Наконец, из этого следует, что в ходе сравнений одного и того же рефлекса с различными другими рефлексами он периодически будет получать различные векторы. Например, вектор рефлекса *a* при сравнении с рефлексом вроде *e* будет короче, чем вектор того же *a* при сравнении с дифтонгом *ei*. Однако, нам не кажется, что это является сколько-либо слабой стороной подхода, поскольку в конечном счёте, мы оперируем не самими векторами, а значениями, полученными через Евклидово расстояние между парами векторов.

3.4. Матрицы томов

После завершения попарных сравнений рефлексов появилась возможность перейти к вычислению итоговых расстояний между диалектами. Для этого была осуществлена обработка каждой из семи таблиц и внутри таблицы – каждой карты.

Для обработки всей таблицы t нам потребуется не только обработать каждую карту в ней, получив в результате матрицы (M_1, M_2, … M_{n-1}, M_n), описанные в следующем абзаце, где n – количество карт в таблице, но и создать матрицу V_t, в которой будут храниться данные о количестве произведённых сравнений.

Обработка карты k представляет собой построение квадратной матрицы M_k размером 780*780, значения которой сначала приравниваются к нулю, а затем заполняются следующим образом: для элемента с индексом (i, j) берём два множества: I – рефлексы, представленные в населённом пункте с номером i, и J – рефлексы в населённом пункте j; если оба множества оказались не пустыми, прямым произведением получаем множество пар рефлексов, между которыми надо произвести сравнение, в качестве значения $M_k(i, j)$ записываем среднее арифметическое набора из $|I| \cdot |J|$ результатов сравнений, значение $V_t(i, j)$ увеличиваем на единицу.

Далее полученные n матриц складываются, получаемая матрица поэлементно делится на матрицу V_t. Иными словами, задаётся матрица T_t такая, что $T_t(i,j) = (\sum_{k=1}^{n} M_k(i,j)) \,/\, V_t(i,j)$, это позволяет нивелировать эффект от различающегося количества карт, по которым проводились сравнения между различными парами диалектов. Матрица T_t – матрица тома, представляет собой матрицу с информацией о языковом расстоянии между каждой из пар диалектов, полученной посредством сравнения рефлексации конкретной фонемы в различных лексемах.

3.5. Итоговая матрица

При попытке сложения матриц томов с целью создания диалектной классификации, основывающейся на вкладе рефлексом всех исследуемых континуантов гласных, возникает существенная методологическая сложность, связанная с количеством карт в разных томах. Так, таблица, отражающая судьбу *е, содержит 73 карты, в то время как в таблице, посвящённой *ъ, содержится 24 карты. Вариант с простым суммированием 7 матриц (T_1, T_2, … T_6, T_7) имеет очевидную проблему: оба рефлекса имеют равную лингвистическую значимость, но каждая отдельная карта на *е будет весить втрое меньше, чем карта на *ъ. Альтернативный вариант предполагает перед сложением умножить каждую из 7 матриц T на количество карт в соответствующей таблице, но в таком случае возникает иная проблема – несмотря на равный вес каждой карты, непропорционально увеличивается вклад рефлексов с большим количеством карт. В нашем примере это приведёт к тому, что совокупный вклад всех рефлексов *е в дифференциацию исследуемых диалектов становится в три раза больше, чем вклад рефлексов *ъ.

Нами был выбран компромиссный вариант, при котором перед сложением каждая матрица умножалась на квадратный корень отношения наибольшего количества карт среди всех таблиц к количеству карт в данной таблице. Так, например, матрица *ъ будет домножена на $\sqrt{73 \,/\, 24} \approx 1{,}744$, в результате чего вклад каждой отдельной карты, посвящённой *ъ, во столько же раз превосходил вклад карты на *е, насколько совокупный вклад всех карт на *е превосходил аналогичный вклад всех карт на *ъ.

3.6. Кластеризация

Полученная квадратная матрица, отражающая языковые расстояния между каждой из пар диалектов, служит основой для проведения кластерного анализа. В настоящей работе была применена иерархическая кластеризация методом Уорда. Это агломеративный алгоритм, нацеленный на обнаружение таких объединений,

при которых будут получаться кластеры с минимальным приростом дисперсии. Такой подход релевантен для лингвистических исследований, так как позволяет выявлять естественные группировки диалектов на основе объективных количественных показателей их сходства и различия.

3.7. Полигоны

Для удобства визуализации мы приняли решение построить диаграмму Вороного, тем самым получив для каждой из пар координат, соответствующей одному из 780 населённых пунктов, полигон. Такие полигоны геометрически точно ограничивают ту область пространства, для которой соответствующий ей населённый пункт является ближайшим из представленных. Внешние полигоны затем были дополнительно ограничены альфа-формой, описанной вокруг множества всех точек, чтобы избежать проблемы потенциально бесконечных полигонов. Стоит заметить, что иногда полигоны заметно не согласуются с границами государств. Например, полигон, соответствующий пункту 282 в северной Богемии ощутимо вклинивается на территорию польской Силезии. Кроме того, несмотря на то, что в Австрии представлены только 6 пунктов, 3 на востоке в Бургенланде и 3 на юге в Каринтии, соответствующие им полигоны, а также полигоны южных пунктов Чехии и западных пунктов Словакии занимают непропорционально большие территории Австрии.

3.8. Цвета

Для визуального представления близостей между диалектными группами матрицы расстояний были преобразованы в синтетические наборы n-мерных векторов с помощью метода многомерного шкалирования. Метод гарантирует получение такого набора, при котором расстояния между векторами в наборе минимально отличаются от расстояний в исходной матрице. В рамках настоящего исследования многомерное шкалирование применяется для получения трёхмерных векторов с дальнейшим преобразованием в пространство RGB цветов. Цвета кластеров определяются как среднее цветов всех точек, попавших в кластер. Таким образом, чем меньше цветовое расстояние между двумя кластерами, тем больше в среднем похожи диалекты, отнесённые к этим кластерам.

4 Анализ

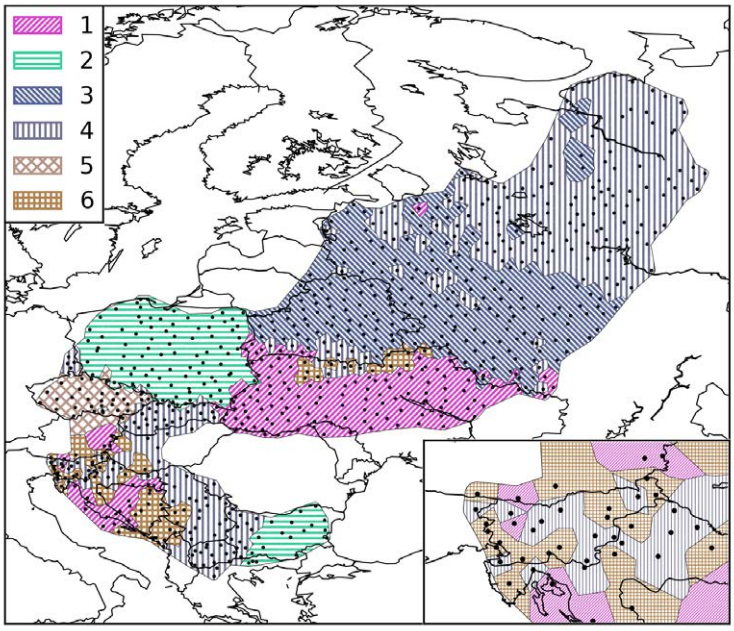
Ниже представлены карты, полученные по данным отдельных томов ОЛА. Опытным путём было обнаружено, что наиболее адекватной (то есть, соответствующей представлениям традиционной компаративистики) кластеризация выходила при делении на 6 кластеров.

Рефлексы *ě (карты *květъ ‘цветок’, *lěsъ ‘лес’, *světъ ‘свет, мир’ и т.д.) разделили славянские говоры на следующие кластеры: 1) идиомы с переходом *ě > *i*, преимущественно украинские и икавские, но также некоторые словенские и один русский; 2) идиомы с рефлексацией *ě > *e* / *a* в результате лехитской перегласовки (*biały*, *śnieg*) и болгарского перехода *ě > *a* под ударением перед твёрдым согласным (*бял*, *сняг*); 3) белорусские и русские идиомы с рефлексацией *ě > *e*, также редукцией безударного *e* (яканье, иканье); 4) идиомы, в которых *ě всегда отражается как *e*; 5) чешские говоры с рефлексацией *i:* / *(j)e* ~ *(ń)e* в зависимости от долготы / краткости ятя (*víra*, но *věřit*); 6) дифтонговые рефлексы в сербохорватских, словенских и украинских говорах – *i̯e*, *ei̯*, *ii̯e* и т.д.

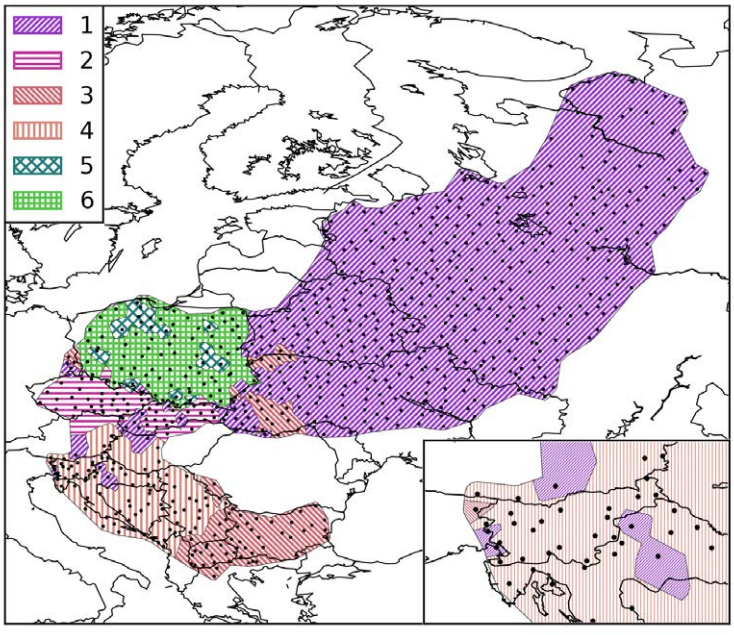
Рефлексы *ę (*ęzykъ ‘язык’, *zajęcь ‘заяц’, *ęčьmy ‘ячмень’ и т.д.) дали следующее деление: 1) идиомы с рефлексацией *ę > *a* или *ą*; 2) чешские и словацкие говоры с рефлексацией *ę > *a* / *e* (в результате разных процессов); 3) идиомы с рефлексацией *ę > *e* и утратой долготной корреляции (болгарские, македонские, часть сербских, нижнелужицкие), а также украинские говоры с рефлексацией *a* / *e* в зависимости от ударения; 4) идиомы с рефлексацией *ę > *e* и сохранением долготной корреляции; 5) кашубские и польские говоры с частично неносовыми рефлексами; 6) польские говоры с преимущественно носовыми рефлексами.

Рефлексы *ǫ (*ǫsъ ‘ус’, *ǫzъkъ(jь) ‘узкий’, *ǫsenica ‘гусеница’ и т.д.) дали следующие кластеры: 1) словенские говоры с переходом *ǫ > *o*, а также чешские говоры с дифтонгизацией и монофтонгизацией *ǫ: > *u:* > *ou̯* > *o:*; 2) чешские говоры с рефлексацией *ǫ > *u* и дифтонгизацией *u:* > *ou̯*; 3) идиомы с рефлексацией *ǫ > *u* и утратой долготной корреляции; 4) идиомы с рефлексацией *ǫ > *u* и сохранением долготной корреляции; 5) польские идиомы с сохранением носовости в том или ином виде; 6) идиомы, в которых *ǫ дал неносовой или преимущественно неносовой рефлекс, отличающийся от *u* и *o*.

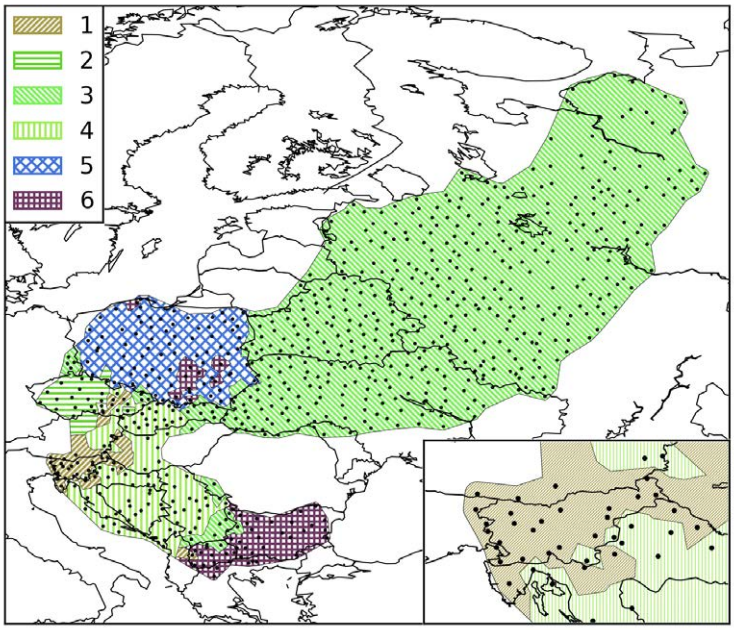
Рефлексы *ъ (*tъ ‘тот’, *sъnъ ‘сон’, *dъžžь ‘дождь’ и т.д.) дали следующую кластеризацию: 1) идиомы с рефлексацией *ъ > *o*, перед *j* – *o* или ы; 2) идиомы с рефлексацией *ъ > *o*, перед *j* – ы; 3) идиомы с рефлексацией *ъ > *o*, перед *j* – *y*; 4) идиомы с рефлексацией *ъ > *e*, перед *j* – *i*; 5) идиомы с рефлексацией *ъ > *a*, перед *j* – *i*; 6) идиомы с рефлексацией *ъ > *ə* или *o,* перед *j* – *i*.



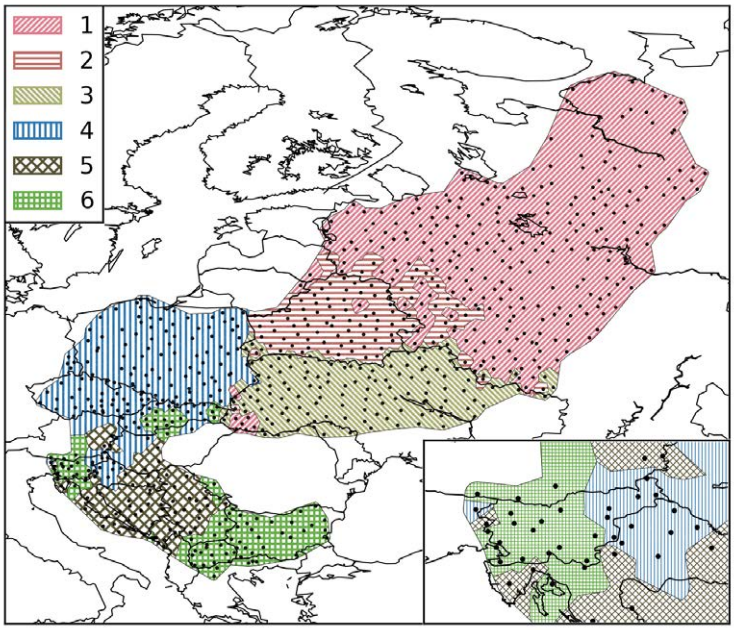
КАРТА 2: Рефлексы *ě



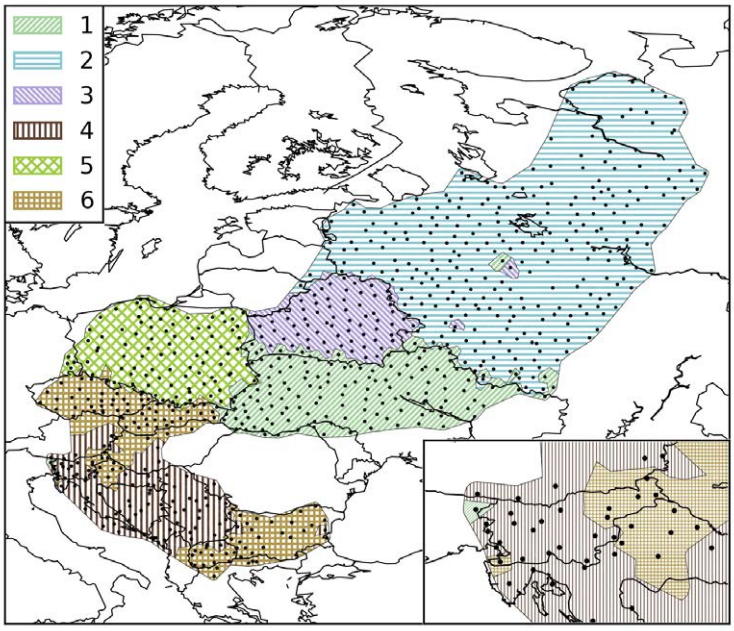
КАРТА 3: Рефлексы *ę



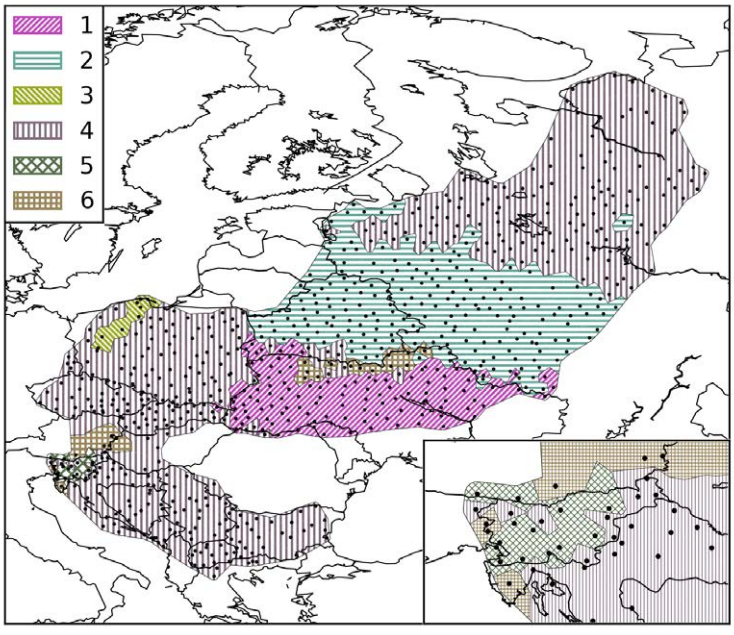
КАРТА 4: Рефлексы *ǫ



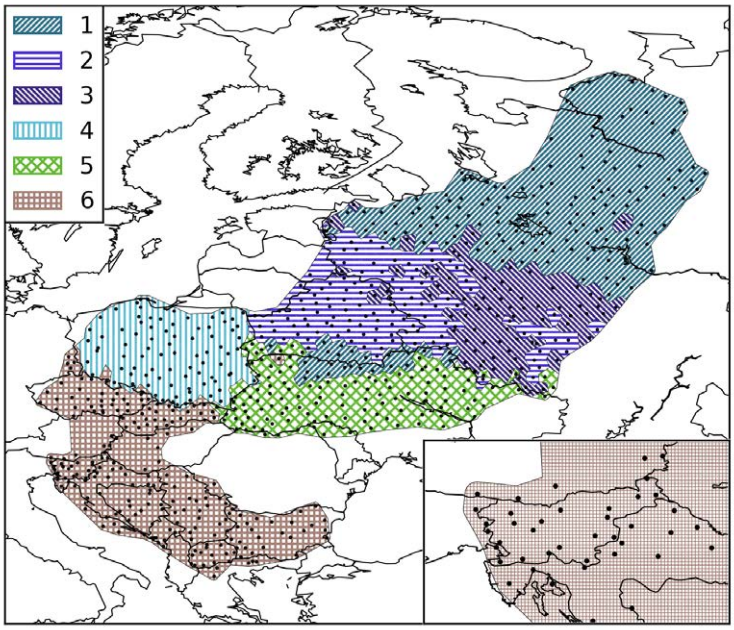
КАРТА 5: Рефлексы *ъ



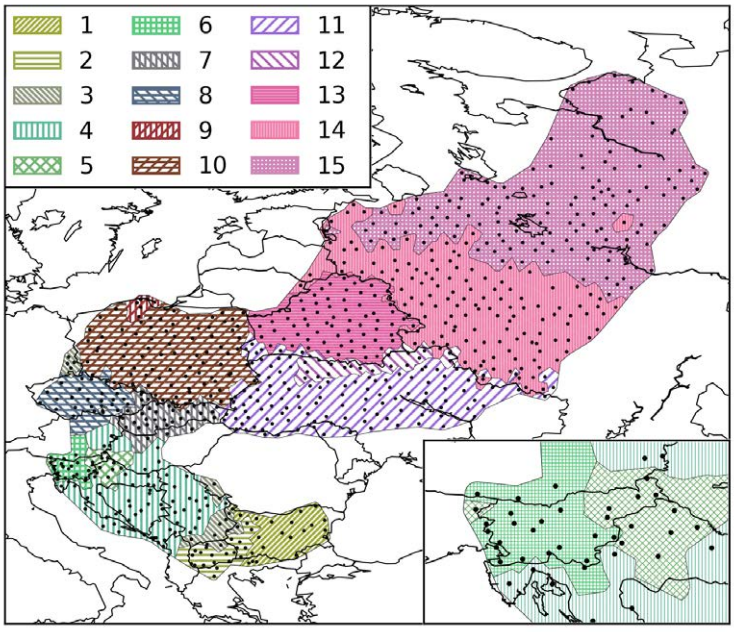
КАРТА 6: Рефлексы *ь



КАРТА 7: Рефлексы *о



КАРТА 8: Рефлексы *e



КАРТА 9: Деление славянских говоров на 15 кластеров на основе всего обработанного материала

Рефлексы *ь (*pьnjь ‘пень’, *dьnь ‘день’, *dьnьsь ‘сегодня’ и т.д.) рисуют следующую картину: 1) идиомы с рефлексацией *ь > *e*, перед *j* – *i* / *y* / 0; 2) идиомы с рефлексацией *ь > *e*, перед *j* – *e* / 0; 3) идиомы с рефлексацией *ь > *e*, перед *j* – *i* / *ы* / 0; 4) идиомы с рефлексацией *ь > *a* или *ə*, перед *j* – *i*; 5) идиомы с рефлексацией *ь > *e*, перед *j* – *i* / *y*[10]; 6) идиомы с рефлексацией *ь > *e*, перед *j* – *i*.

Рефлексы *o (*onъ ‘он’, *ognь ‘огонь’, *osmь ‘восемь’ и т.д.) дают следующие кластеры: 1) украинская рефлексация *o > *o* / *i* (в новом закрытом слоге); 2) рефлексация *o > *o* с накладывающимся сверху аканьем; 3) говоры с дифтонгизацией *o* > *u̯o* в части случаев; 4) идиомы с *o* в большинстве позиций; 5) словенские говоры с *u:* и разнообразными “широкими” дифтонговыми рефлексами в части позиций – *u̯a, u̯ɔ:, u̯:ə, a:u̯* и т.д.; 6) идиомы с “узкими” дифтонговыми рефлексами в части позиций – *u̯o, u̯y* и т.д.

Рефлексы *e (*eže ‘ёж’, *edla ‘ель’, *edinъ ‘один’ и т.д.) показывают следующее деление: 1) идиомы с рефлексацией *e > *e* / *o* (в начале слова и под ударением перед твёрдыми согласными); 2) идиомы с рефлексацией *e > *e* / *o* (в начале слова и под ударением перед твёрдыми согласными) и наложившимся сверху яканьем; 3) идиомы с рефлексацией *e > *e* / *o* (в начале слова и под ударением перед твёрдыми согласными) и наложившимся сверху иканьем; 4) идиомы с рефлексацией *e > *e* / *o* по лехитской перегласовке; 5) идиомы с рефлексацией *e > *e* / *o* (в начале слова) / i (в новом закрытом слоге); 6) идиомы с сохранением *e* в большинстве позиций.

Наложение всего обработанного материала на карту показало, что наилучший результат (не слишком крупное деление, но и не слишком дробное), можно увидеть на карте с 15 кластерами (карта 9).

Распределение населённых пунктов между кластерами выглядит следующим образом:

1. (восточноболгарские говоры): 113а, 114, 116, 120, 124, 127–131, 133–138, 141, 143, 144, 850, 851.
2. (македонские и западноболгарские говоры): 90, 92, 94, 96, 97, 99–105, 110, 111, 119, 121–123, 125, 126, 132.
3. (торлакские и лужицкие говоры): 84–87, 95, 98, 115, 117, 118, 168, 169, 234–237.

[10] *y* вместо *i* в формах вроде *czyj*, *szyja* в польских и лужицких говорах является результатом отвердения шипящих, однако в связи с малым количеством карт на *ь этого вторичного изменения хватило, чтобы отделить польские и лужицкие говоры от чешских и словацких, в которых изначальная рефлексация *ь была идентична.

4. (сербохорватские говоры): 22, 24, 25, 27, 33, 36–63, 65–83, 88, 146a, 147a, 148a, 150–153.
5. (“западнословенские” и один кайкавский говор): 2–17, 26, 146, 147, 148.
6. (“восточнословенские” и часть кайкавских говоров): 1, 18–21, 28–32, 35, 149.
7. (словацкие и часть чешских говоров): 154–156, 197, 200, 202–206, 208–232, 299.
8. (чешские говоры): 175–196, 198, 199, 201.
9. (кашубские говоры): 241–245.
10. (польские говоры): 238–240, 246–252, 254–298, 300–306, 308–326.
11. (украинские говоры): 171, 233, 361, 362, 372–374, 382, 401–409, 412–416, 419–422, 427–434, 443, 448–524, 836, 843, 847, 849.
12. (полесские говоры): 383, 410, 411, 417, 418, 423–426, 435–442, 444, 446, 447.
13. (белорусские говоры): 328–360, 363–371, 375–381, 384–400.
14. (южнорусские говоры): 445, 525–527, 560, 562, 579, 580, 605, 651–654, 656, 658, 660, 662, 670, 674, 675, 678–680, 692–698, 710–716, 725–730, 732–734, 746–754, 758–765, 769, 771–807, 809–835, 837–841, 844–846, 848.
15. (севернорусские говоры): 528–533, 535–537, 540–549, 551, 552, 554–557, 563–565, 567–578, 581–593, 596, 597, 600, 602–604, 606–614, 616–628, 631–644, 646–648, 655, 657, 659, 661, 663-668, 671, 673, 681–691, 699–707, 709, 718–720, 722–724, 738, 742–745, 756, 757, 770.

Полученное распределение по большей части соответствует представлениям компаративистики о делении славянского языкового континуума[11]. Нет ошибок в проведении границ между восточно-, западно- и южнославянскими пунктами, например, пункт 233 в Словакии верно аттрибутирован как украинский. Восточноболгарские идиомы отделены от западноболгарских и македонских по ятевой границе. Алгоритм правильно отделил кашубские говоры от польских, а также польские от чешских и словацких. Также он совершенно верно определил украинские пункты на территории России (836, 843, 847, 849 с рефлексацией *ě > *i*, оканьем, а также *i* в новом закрытом слоге) и русский на территории Украины (445 с *ě > *e*, переходом *e* > *o* под ударением (*мёд*), аканьем и, что самое важное, яканьем). Справедливо, на наш взгляд, западнополесские говоры на территории Белоруссии (361, 362, 372–374, 382) были отнесены к украинским (*ě > *i*, оканье, *i* в новом закрытом слоге). Правильно аттрибутирован говор 670 в составе Чухломского акающего острова.

[11] Снова напомним, что полученная карта отражает современное состояние славянских языков (включены переселенческие говоры, в том числе XX века).

Конечно, есть и недостатки. Наибольшим является ложный кластер, объединяющий торлакские говоры с лужицкими. Также не вполне понятно, насколько обоснованно включение в него пунктов 95, 98, 115, 117, 118 (из-за рефлексации *ǫ > *u*) и переселенческих 168 и 169 (из-за рефлексации *ъ, *ь > *ə* / *a*), кажется, данных одних только гласных для решения этого вопроса недостаточно. Необоснованно деление словенских говоров на две части и объединение их с кайкавскими по отдельности. Некорректна также полученная чешско-словацкая граница.

Перечисленные недостатки вызваны тем, что классификация основывается исключительно на данных вокализма. В то же время её достоинства позволяют полагать, что, когда выйдут тома, посвящённые консонантизму, удастся построить более надёжную и исчерпывающую классификацию славянских говоров. Кроме того, в нашей работе мы отказались от арбитрарного подхода к выбору весов для рефлексов разных гласных, так как не хотели вновь возвращаться к стандартным качественным подходам в лингвогеографии, при которых классификация языков может существенно меняться в зависимости от того, что конкретный исследователь считать более важным.

Второй путь дальнейших исследований, по которому авторы данной работы надеются направиться в ближайшее время, заключается в создании классификации говоров ОЛА на основе диахронных фонетических изменений, а не близости рефлексов в современных идиомах.

Статья основана на научных данных из опубликованных и общедоступных источников (в первую очередь Общеславянского лингвистического атласа), список которых приведен в разделе « библиография».

БИБЛИОГРАФИЯ

Васильев, Михаил Е., Саенко, Михаил Н. 2020. Анализ топологии и оценка точности лексикостатистических классификаций (на примере славянских языков). *Вопросы языкового родства*, 18/3–4. 130–157. [Vasil'jev, Mikhail Je., Saenko, Mikhail N. 2020. Analiz topologii i otsenka tochnosti leksikostatisticheskikh klassifikatsii (na primere slavianskikh iazykov). *Voprosy iazykovogo rodstva*, 18/3–4. 130–157.]

Кузьмина, Анастасия С., Манусов, Арсений В. 2024. Диалектометрический подход к диалектной классификации восточнославянских языков на материале сборника «Восточнославянские изоглоссы». *Вопросы языкового родства*, 22/3–4. 342–366. [Kuz'mina, Anastasiia S., Manusov, Arsenii V. 2024. Dialektometricheskii podkhod k dialektnoi klassifikatsii vostochnoslavianskikh iazykov na materiale sbornika «Vostochnoslavianskije izoglossy». *Voprosy iazykovogo rodstva*, 22/3–4. 342–366.]

Марченко, Игорь А., Ронько, Роман В. 2025. Диалектные различия между востоком и западом на материале данных Диалектологического атласа русского языка: результаты многомерного шкалирования. *Исследования по славянской диалектологии* 25. 236–259. [Marchenko, Igor' A., Ron'ko, Roman V. 2025. Dialektnyje razlichiia mezhdu vostokom i zapadom na materiale dannykh Dialektologicheskogo atlasa russkogo iazyka: rezul'taty mnogomernogo shkalirovaniia. *Issledovaniia po slavianskoi dialektologii* 25. 236–259.]

ОЛА = *Общеславянский лингвистический атлас. Серия фонетико-грамматическая*, 1–9. 1988–2019. Београд, Москва, Wrocław, Warszawa, Kraków, Zagreb, Скопје, Мінск, Praha, Bratislava, Санкт-Петербург. [OLA = *Obshcheslavianskii lingvisticheskii atlas. Seriia fonetiko-grammaticheskaia*, 1–9. 1988–2019. Beograd, Moskva, Wrocław, Warszawa, Kraków, Zagreb, Skopjje, Mīnsk, Praha, Bratislava, Sankt-Peterburg.]

Прохоров, Валентин А. 1973. *Вся Воронежская земля. Краткий историко-топонимический словарь*, Воронеж. [Prokhorov, Valentin A. 1973. *Vsia Voronezhskaia zemlia. Kratkii istoriko-toponimicheskii slovar'*, Voronezh.]

Пшеничнова, Надежда Н. 1996. *Типология русских говоров*, Москва. [Pshenichnova, Nadezhda N. 1996. *Tipologiia russkikh govorov*, Moskva.]

Старостин, Сергей А. 2007. Сравнительно-историческое языкознание и лексикостатистика. В: С. А. Старостин. *Труды по языкознанию*. М.: Языки славянских культур. 407–447. [Starostin, Sergei A. 2007. Sravnitel'no-istoricheskoje iazykoznanije i leksikostatistika. V: S. A. Starostin. *Trudy po iazykoznaniiu*. M.: Iazyki slavianskikh kul'tur. 407–447.]

Afanasev, Ilia. 2023. Cipher, transform, get lost: an anti-transparent system for distance measurement in East Slavic lects. *Journal of Language Relationship*, 21/3–4. 159–177.

Houtzagers, Peter, Nerbonne, John, Prokić Jelena. 2010. Quantitative and Traditional Classifications of Bulgarian Dialects Compared. *Scando-Slavica* 56. 29–54.

Kenda-Jež, Karmen. 2012. Slovenska narečja v Slovanskem lingvističnem atlasu (OLA). *Slavistica Vilnensis*, 57/2. 57–76.

Kushniarevich, Alena et al. 2015. Genetic heritage of the Balto-Slavic speaking populations: Asynthesis of autosomal, mitochondrial and Y-chromosomal data. *PLoS ONE*, 10/9.

Mańczak, Witold. 2006. *Pochodzenie języka staro-cerkiewno-słowiańskiego a kodeks Zografski*. Warszawa: Zakład Graficzny Uniwersytetu Warszawskiego.

McMahon, April, McMahon, Robert, 2005. *Language Classification by Numbers*. Oxford.

Saenko, Mikhail N. 2020. Taxonomy of Slavic Languages, History of the. In: Marc L. Greenberg, ed. *Encyclopedia of Slavic Languages and Linguistics Online*, Leiden.

Šekli, Matej. 2018. *Tipologija lingvogenez slovanskih jezikov*. Ljubljana.

Povzetek

Klasifikacija slovanskih jezikov na podlagi Slovanskega lingvističnega atlasa: prvi poskus

Prispevek predlaga klasifikacijo slovanskih jezikov, ki temelji na objektivni dialektometrični analizi gradiva Slovanskega lingvističnega atlasa (OLA). Za

razliko od tradicionalnih klasifikacij, ki se opirajo predvsem na knjižne jezike in so pod vplivom zunajjezikovnih dejavnikov, se ta metoda osredotoča na narečni kontinuum.

Delo temelji na podatkih iz šestih zvezkov fonetično-slovnične serije OLA, ki vsebujejo informacije o refleksih sedmih praslovanskih samoglasnikov (*ě, *ę, *ǫ, *ъ, *ь, *o, *e) v 780 krajih. Za vsak refleks so bile določene jezikovne značilnosti (mesto tvorbe, dolžina, labializacija, nazalizacija) in predstavljene kot vektorji. Ključna faza je bila parna primerjava refleksov med vsemi točkami z izračunom evklidske razdalje, pri čemer so bile upoštevane strukturne razlike (na primer med monoftongi in diftongi). Na podlagi konstruirane matrice jezikovnih razdalj je bilo uporabljeno hierarhično združevanje po Wardovi metodi.

Analiza je na podlagi obravnavanega gradiva pokazala optimalno delitev slovanskega območja na 15 skupin. Algoritem je potrdil tradicionalne koncepte slovanske dialektologije, saj je pravilno razdelil vzhodno-, zahodno- in južnoslovansko območje ter pravilno pripisal točke zunaj državnih meja. Obenem pa so bile razkrite tudi pomanjkljivosti, povezane z zanašanjem zgolj na vokalizem, med katerimi je bila najbolj presenetljiva lažna združitev torlaških in lužiških narečij.

Študija prikazuje potencial kvantitativnih metod za identifikacijo strukture narečnega kontinuuma, brez subjektivnih ocen. Nastala klasifikacija služi kot pomembna referenčna točka, vendar ne nadomešča tradicionalnih pristopov.

A Classification of the Slavic Languages based on the material of the Slavic Linguistic Atlas: A Trial Run

This study proposes a classification of Slavic languages based on an objective dialectometric analysis of the materials of the Slavic Linguistic Atlas. Unlike traditional classifications, which rely primarily on literary languages and are influenced by extralinguistic factors, this method focuses on the dialect continuum.

The work is based on data from 6 volumes of the phonetic-grammatical series of the Slavic Linguistic Atlas, containing information on the reflexes of 7 Proto-Slavic vowels (*ě, *ę, *ǫ, *ъ, *ь, *o, *e) in 780 localities. For each reflex, linguistic characteristics (height, backness, length, rounding, nasalization) were determined and presented as vectors. The key stage was a pairwise comparison of reflexes between all locations with the calculation of the Euclidean distance, taking into account structural differences (for example,

between monophthongs and diphthongs). Based on the constructed matrix of linguistic distances, hierarchical clustering by Ward's method was applied.

The analysis revealed that, for the material considered, the optimal division of the Slavic area is into 15 clusters. The algorithm confirmed established views in Slavic dialectology, as it correctly separated the East, West, and South Slavic areas and accurately attributed locations outside state borders. However, shortcomings associated with relying solely on vocalism were also identified, the most striking of which was the erroneous unification of the Torlak and Sorbian dialects.

The study demonstrates the potential of quantitative methods to reveal the structure of a dialect continuum, free from subjective assessments. The resulting classification serves as an important benchmark but does not replace traditional approaches.