А. С. Кузьмина, А. В. Манусов†*
* Высшая школа экономики, Москва; askuzmina5487@gmail.com
† Высшая школа экономики, Москва; arseniymanusov@gmail.com

Диалектометрический подход к диалектной классификации восточнославянских языков на материале сборника «Восточнославянские изоглоссы»

В тексте работы предложен новый диалектометрический подход к членению восточнославянских языков. Наша диалектометрия основана на материале сборника статей «Восточнославянские изоглоссы» (ВСИ-1 1995; ВСИ-2 1998; ВСИ-3 2000; ВСИ-4 2006), который представляет собой обобщение данных атласов восточнославянских языков (АУМ, ДАБМ, ДАРЯ). Карты ВСИ были перенесены нами в электронный формат, при этом каждый признак, представленный на изначальной карте, хранится как отдельное изображение. Эта процедура позволила построить матрицу признаков, содержащихся в восточнославянских атласах. Был использован метод *k*-средних для построения кластеризации. Получены новые варианты диалектных членений ВСЯ, для ареалов выявлены значимые диалектные черты, которые являются дифференциальными. Данное диалектное членение сравнивается с уже существующими. Материалы ВСИ содержат однородные данные для всех трёх языков, это позволяет работать с территорией атласов, как с единым языковым ареалом, благодаря чему появляется возможность исследовать ВСЯ не только дискретно, но и континуально. В статье предложен анализ трансграничных говоров представленного диалектного членения.

Ключевые слова: диалектометрия; количественные методы в лингвистике; восточнославянские языки; славянские языки; изоглоссы; автоматическая классификация языков; диалектология; лингвогеография.

Введение

Лингвистическая география, как самостоятельная область лингвистического знания, начала формироваться в сер. XIX — начале XX века. Толчком к её развитию послужили работы учёных-лингвистов, сосредоточивших свои усилия на составлении национальных языковых атласов: Лингвистического атласа Франции 1912 г. и Языкового атласа Германской империи 1926–1954 гг. Эти исследования, вместе с развивающимися в сер. XIX в. теорией лингвистической непрерывности А. Пихте (языки не существуют изолированно, а образуют непрерывный континуум) и теорией волн Г. Шухардта и И. Шмидта (языковые инновации распространяются из центра их возникновения к периферии подобно волнам, затухающим по мере удаления от источника), подчеркнули необходимость систематического изучения географического распределения языковых явлений.

Изначально в лингвогеографии использовался *качественный метод* сравнения, в котором решение о том или ином диалектном членении принимал человек, самостоятельно оценивающий вклад и значимость тех или иных признаков при создании классификации. При таком методе обычно бралось ограниченное число изоглосс, и достаточно часто привлекались экстралингвистические данные (этнический состав, политиче-

ские границы, миграционные волны). Для выделения говоров традиционно используются результаты единообразных анкетирований, проводимых в населённых пунктах. Анкетирование проводится по заранее разработанному комплекту вопросов, полученные ответы визуализируются на картах. Первая попытка применить данный метод для всего восточнославянского ареала — работа московской диалектологической комиссии под руководством Н. Н. Дурново, и, как результат, опубликованная ими «Диалектологическая карта русского языка в Европе 1915 года» (Дурново и др. 1915). Эта карта принципиально отличается от предыдущих диалектных карт русского языка, так как разделение языка на более мелкие единицы осуществлялось исключительно на основе *лингвистических* принципов, в отличие от прежних подходов, учитывающих совокупность этнографических особенностей конкретного региона.

В 1965 году К. Ф. Захарова и В. Г. Орлова (Захарова & Орлова 1970) опубликовали карту диалектного членения русского языка (ДАРЯ Вып. III. Часть 2. Карта II) на материале сведений, собранных по программе собирания сведений (Аванесов 1947) для создания ДАРЯ. Это позволило получить единообразную информацию обо всех населённых пунктах, в которых проводилось анкетирование. Благодаря плотной сетке деревень, а также участию в сборе сведений специалистов-лингвистов удалось собрать подробные сведения о русских говорах. К. Ф. Захарова и В. Г. Орлова выделили 24 группы говоров на территории говоров русского языка первичного образования.

Другой метод, получивший развитие в конце XX века — *метод количественный* (диалектометрия). Он основан на применении компьютерных алгоритмов к анализу диалектных черт. Вначале он сводился к попарному сравнению признаков из ближайших населенных пунктов, но постепенно усложнился, позволив осуществлять сравнение между N-м количеством деревень и сёл. Таким образом, в лингвогеографии стало возможным рассмотрение всего собранного диалектного материала, ограничиваясь только мощностями компьютера и объёмом оперативной памяти.

Примерами применения количественных методов для восточнославянских языков могут быть 2 исследования: «Структурно-типологическая классификация русских говоров» Н. Н. Пшеничновой (Пшеничнова 1996) и «Классификация русских говоров с применением многомерного шкалирования» И. А. Марченко (Marchenko 2023).

«Структурно-типологическая классификация русских говоров» Н. Н. Пшеничновой (Пшеничнова 1996; ДАРЯ Вып. III. Часть 2. Карта III) — это очень детальное, подробное членение, опирающееся на анализ обширного массива данных: в его основу легли все диалектологические материалы ДАРЯ за исключением статистических карт. В отличие от предшественников, выделяющих ареалы по пучкам изоглосс, Н. Н. Пшеничнова предложила структурно-типологическую группировку говоров, построенную на основании их полных описаний по всем языковым уровням. Для группировки говоров используется понятие таксонометрического отношения, используемого в биологии. При таком подходе связь между говорами считается сильной, если в сравниваемых говорах есть редкие признаки или отсутствуют признаки, распространённые в других областях широко (Пшеничнова 1996: 10–11).

В работе 2023 г. И. А. Марченко (Marchenko 2023) предлагает для построения диалектных членений русского языка воспользоваться методом многомерного шкалирования. Данный метод предназначен для проекции точек на пространство меньшей размерности таким образом, чтобы различия между новыми расстояниями по сравнению с первоначальными были как можно меньше (Marchenko 2023: 11). В этой работе метод позволил осуществить переход от пространства, размерность которого определена общим количеством признаков к трёхмерному пространству, однозначно задающему цвет

каждого набора данных. Для построения использовались материалы ДАРЯ: матрицы признаков вида *населенные пункты* × *диалектные признаки*.

В этой работе мы предлагаем новый диалектометрический подход к членению восточнославянских языков. Нам неизвестны работы, посвященные диалектному членению всего восточнославянского ареала на основании *количественных* методов, поэтому эта работа позволит впервые определить и уточнить положение приграничных говоров в восточнославянском диалектном континууме с помощью компьютерных методов.

В разделе 1 данной статьи содержится информация об используемых материалах и о созданном нами инструменте по работе с диалектными картами восточнославянского ареала. В разделе 2 описаны методы работы с картами: создание матрицы признаков, кластеризация матрицы, придание кластерам цветов и визуализация результатов на карте. Также содержится описание метрики, используемой в разделе 3 для оценки значимости диалектных признаков в полученной кластеризации. В третьем разделе дается подробное описание полученных кластеризаций и сравнение их с существующими диалектными членениями.

За границей нашего рассмотрения оказались кластеризации по языковым разделам, так как в силу малого объема карт генерации становятся достаточно нестабильными. Также не было учтено несбалансированное количество признаков: синтаксических карт было втрое меньше, чем карт по другим разделам.

Публикация подготовлена в ходе проведения исследования № 25-00-019 «Корпусное и ареальное изучение восточнославянских диалектов» в рамках Программы «Научный фонд Национального исследовательского университета “Высшая школа экономики” (НИУ ВШЭ)».

1. Материалы и электронный ресурс

Кластеризации в работе построены на картах из сборника статей «Восточнославянские изоглоссы» (ВСИ-1 1995; ВСИ-2 1998; ВСИ-3 2000; ВСИ-4 2006), описывающего говоры восточнославянского ареала на материале ДАРЯ (*Диалектологический атлас русского языка*) (*ДАРЯ* 1986-2004), *ДАБМ* (*Дыялекталагічны атлас беларускай мовы*) (*ДАБМ* 1963), АУМ (Атлас української мови) (АУМ 1984-2001). Помимо рекартирования упомянутых национальных атласов также привлекались региональные атласы, карты Общеславянского лингвистического и Общекарпатского диалектологического атласов. Таким образом, в сборник ВСИ вошло много диалектных карт, значительно отличных от представленных в национальных атласах.

Статьи в 4-х томах ВСИ посвящены явлениям из области фонетики, морфологии, синтаксиса и лексики. Мы считаем, что «Восточнославянские изоглоссы», благодаря отображению в них большого количества диалектных явлений (67), являются подходящим материалом для нашего исследования. Однако работа над сборником была приостановлена в конце 2000-х гг. и несмотря на довольно большое количество картографированных диалектных явлений, некоторые весьма существенные для диалектного членения ВСЯ явления не вошли в сборник и, следовательно, в нашу работу.

Преобразовать исходные карты из «Восточнославянских изоглосс» вручную или автоматически в матричный формат для дальнейшего компьютерного анализа не представляется возможным, в силу отсутствия информации о населённых пунктах, учтённых при создании карт, а также в силу различных обобщений и разной степени детализации отображения диалектных черт. Поэтому было решено перерисовать все карты в электронный формат и из них извлекать сведения о диалектных признаках.

Для этого вначале был создан базовый слой 984 на 969 пикселей с территорией, отцентрованной по исторической зоне расселения восточных славян. На базовый слой были перенесены основные водоемы: реки, озера, водохранилища; также были созданы 3 подложки (имперские губернии, советские республики, советские области), позволяющие соотнести распределение диалектных черт с определёнными политическими и административными территориями. Отдельно была создана база данных, содержащая информацию о названиях всех карт и о легендах к ним. Карты были перерисованы с помощью программы paint.net, каждая диалектная черта была нарисована на отдельном слое; для визуализации перерисованных карт, а также для возможности наложения одних диалектных черт (слоев) на другие был разработан и опубликован в свободном доступе электронный ресурс (Манусов и др. 2024). На нём содержатся все диалектные карты ВСЯ из 4-х томов «Восточнославянских изоглосс» (ВСИ-1 1995; ВСИ-2 1998; ВСИ-3 2000; ВСИ-4 2006), осуществлен поиск, доступны сортировки по авторам, темам и томам.

2. Методы

Компьютерная обработка нарисованных слоев и создание кластеризаций осуществлялись с помощью языка программирования Python. Вначале слои были преобразованы в матрицу признаков из нулей и единиц (п. 2.1), затем на основе матрицы была построена кластеризация (п. 2.2) и визуализирована на карте (п. 2.3). Полученные кластеры были оценены с помощью метрики FM (п. 2.4).

2.1. Матрица признаков

На основе нарисованных слоев была создана матрица признаков: таблица, в которой каждому пикселю соответствует 1, если диалектная черта встречается в этом пикселе, и 0 — если она отсутствует. Таким образом, в нашей работе для всех диалектных признаков были использованы *равные* веса. Всего на 67 картах представлен 341 ареал и 105 изоглосс. Изоглоссы были дополнительно перерисованы как ареалы, для удобства дальнейшей работы с ними. Маркеры не были использованы для анализа всего восточнославянского континуума в силу трудности локализации признака на карте. Всего получилось 446 слоев: из них посвящены фонетике 113 слоёв, морфонологии и морфологии — 158 слоёв, синтаксису — 39 слоёв, лексике — 136 слоёв. Итого, получилась матрица размером 406782*446 (Схема 1), из которой предварительно исключили все точки, лежащие за пределами исследуемой области (спорадически встречающиеся описания диалектных черт на территориях восточной Польши, Словакии, Эстонии, Латвии, Литвы, Молдавии), а также точки, которые были описаны на малом количестве карт (в частности, были исключены южные территории по линии Подольск-Днепр, Днепр-Чугуев). В дальнейшем исследовании были использованы данные со следующих территорий: территории России (ограничено территориями говоров первичного формирования), Украины (кроме юга) и всей Беларуси.

$$\begin{array}{c} \\ \mathbf{1} \\ \mathbf{2} \\ \cdots \\ \mathbf{406782} \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{1} & \mathbf{2} & \ldots & \mathbf{446} \end{array} \\ \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 1 \end{pmatrix} \end{array}$$

Схема 1. Двухмерная матрица, в которой столбцы соответствуют диалектным признакам, а строки — пикселям на карте.

Важно отметить, что далеко не на всех точках карты есть населенные пункты, и, следовательно, не везде были действительно замечены те или иные явления. В работе мы делаем обобщение: если пиксели находятся внутри некоторого языкового ареала, то все они наделены этим диалектным признаком. Такой подход принципиально отличается от того, как работают с диалектными атласами восточнославянских языков, в которых достоверно известно, какие именно населенные пункты были исследованы.

2.2. Кластеризация

Для кластеризации (упорядочивания данных в группы на основании их близости) был использован метод *k*-средних (*k*-means). Данный метод минимизирует сумму квадратичных отклонений всех точек кластера от его центра (Схема 2):

$$\sum_{j=1}^{k} \sum_{x_i \in C_j} \left(x_i - \mu_j\right)^2$$

Схема 2. Метод *k*-средних,
где j — некоторый кластер, μ_j — центр кластера j, k — количество кластеров, C_j — множество точек, назначенных кластеру j, x_i — некоторая точка кластера j.

Опишем работу алгоритма: вначале выбирается некоторое количество случайных центров, их количество — гиперпараметр, то есть задается вручную. Каждый центр — это набор чисел длины 446, что соответствует количеству признаков в наших данных. После для каждой из 406782 точек алгоритм подбирает такой центр, расстояние к которому будет минимальным, тем самым разбивая изначальный массив данных на кластеры. Затем для получившихся кластеров считаются центры масс, в нашем случае в силу бинарности данных представляющие собой вектора длины 446, каждое из значений которых это доля точек в кластере, в которых представлен признак, по отношению ко всем попавшим в кластер точкам. Первичные случайные центры заменяются на вычисленные центры масс, после чего точки заново распределяются между центрами аналогичным образом. Две операции повторяются, что постепенно приводит к наилучшей аппроксимации и всё меньшему отличию результата каждой новой итерации. Однако у такого метода есть существенные проблемы — он очень чувствителен к первой случайной генерации. Из-за этого каждая новая генерация достаточно сильно может отличаться от предыдущей, что не подходит для нашего исследования. Кроме того, *k*-средних достаточно ресурсоёмкий и при наших размерах данных выполняется медленно. Поэтому мы воспользуемся не классическим методом *k*-средних, а методом глобальных *k*-средних (global *k*-means) (Likas et all. 2003). Он не только сводит разницу в генерациях к минимуму, но и значительно снижает вычислительную нагрузку при относительно небольшой потере качества.

Исходя из описания работы данного алгоритма необходимо сформулировать важную его особенность: выделение новых классификационных единиц не всегда происходит из части одного уже существующего кластера. Новый кластер, может возникнуть на территории сразу нескольких более ранних диалектных единиц. Другими словами, границы кластеров, выделившихся позднее, не будут укладываться в границы более ранних.

Как было сообщено выше в этом пункте статьи, количество кластеров — это гиперпараметр, который мы вводим самостоятельно, а не определяем его автоматически. Мы решили остановиться на двух кластеризациях и описать их подробнее: на 5 и на 25 кластерах.

Изначально мы остановились на числе 50, так как это число приближено к общему количеству групп говоров в источниках, на которое мы опирались. Так, русских говоров

на карте Захарова & Орлова 1970 — 26, украинских говоров — 14 (опирались на упрощенный вариант диалектного членения Довгопол и др. 1977), белорусских говоров — 7 (Аванесаў и др. 1969), что в сумме нам дает 47 ареалов, а также один ареал Брянского угла, который не был отнесен на упомянутых классификациях ни в одну группу. Однако кластеризация на 50 кластеров получается крайне нестабильной: при кластеризации на большое количество единиц разница между кластерами сокращается, из-за чего стабильность изображения постепенно падает и вклад каждого отдельного диалектного признака в созданный кластер становится ниже. В связи с чем было принято решение построить кластеризацию на 25 кластеров — половину от 50-ти. При кластеризации на 25 генерации становятся достаточно стабильными, при этом ареалы являются достаточно информативными в лингвогеографическом аспекте и может быть сопоставлены с уже имеющимися диалектными классификациями (без учёта в них переходных говоров и подгрупп).

Также была выбрана классификация на 5 кластеров — это оптимальное значение, полученное с помощью инструментов анализа данных — «метода локтя»[1] (elbow method).

2.3. Генерация цветов и изображений

Чтобы отразить близость кластеров друг к другу мы воспользовались методом уменьшения размерности ICA (independent component analysis) (Comon 1994; Hyvärinen et al. 2001), алгоритмом FastICA (Hyvärinen & Oja 1997) — анализом независимых компонент, быстро сходящимся к наиболее точному решению. Вручную было установлено необходимое количество компонент — 3 — соответствующее 3 цветам палитры RGB: красному, зеленому, синему.

Далее полученный набор данных независимо по каждой компоненте был нормализован, то есть приведен к интервалу от 0 до 1. В результате мы получили матрицу, в которой каждому пикселю на карте стали соответствовать 3 числа, т. е. их сгенерированный цвет, отражающий отношения расстояний между кластерами. Затем было получено изображение в растровом формате PNG.

2.4. Оценка значимости выделяемых кластеров

Для каждого кластера с помощью метрики FM (Fowlkes–Mallows index) (Схема 3) были определены значимые диалектные черты, которые выделяют данный кластер и отличают его от других. Для данной метрики существенным является количество пикселей, совпавших у кластера и диалектного ареала (TP), а также размеры кластера (FP+TP) и размеры ареала (TP+FN). Значения метрики лежат в диапазоне [0, 1]; чем ближе значение к 1, тем лучше определенный слой на карте объясняет кластер, так как в таком случае пересечение будет наибольшим, а количество пикселей вне пересечения — наименьшим, или вовсе равным нулю, если ареал и кластер полностью совпали.

$$FM = \sqrt{PPV * TPR} = \sqrt{\frac{TP}{TP + FP} * \frac{TP}{TP + FN}}$$

Схема 3. формула нахождения FM,
где TP — пересечение кластера и ареала (true positives), FP — размер кластера без пересечения (false positives), FN — размер ареала без пересечения (false negatives), PPV — отношение пересечения к размеру кластера (positive predictive rate), TPR — отношение пересечения к размеру ареала (true positive rate).

[1] График, на котором отмечена доля объясненной этим кластером дисперсии по оси Y и количество кластеров по оси X.

При попарном сравнении кластеров (п. 3.1.3 Таблицы 3, Таблица 4 и п. 3.2.3 Таблицы 7) подсчитывается разница FM метрики по каждому диалектному признаку. Чем ближе эта разница по модулю к 1, тем сильнее этот диалектный признак отличает один кластер от другого. При определении признаков, наилучшим образом объединяющих два кластера, FM метрики наоборот суммируются. Чем ближе их значение к 2-м, тем лучше черта описывает единство кластеров.

3. Результаты

В результате были получены кластеризации на 5 (п. 3.1) и на 25 (п. 3.2) диалектных кластеров. Черно-белая печать не позволяет в полной мере передать близость кластеров друг к другу, так как превращает трехмерную цветную размерность в одномерную, поэтому мы дополнительно прикрепляем таблицы, отражающие относительную близость количественно. Для каждой кластеризации проводится сравнение с существующими диалектными членениями русского, белорусского и украинского ареалов (п. 3.1.1, п. 3.2.1), построенными качественным методом, для этого было выбраны следующие работы: для русского языка — (Захарова, Орлова 1970), для белорусского языка — (Аванесаў и др. 1969), для украинского языка — (Довгопол и др. 1977). Также строятся изображения, на которых границы наших диалектных членений накладываются на границы ареалов в данных работах (Рисунок 3, Рисунок 5). Предлагается анализ нескольких выделенных кластеров, а именно, определение для них наиболее значимых дифференцирующих диалектных признаков (п. 3.1.2, п. 3.2.2). В конце предлагается континуальный подход к классификации диалектов (п. 3.1.3, п. 3.2.3).

3.1. Кластеризация на 5 кластеров

В результате кластеризации на 5 кластеров и последующего применения алгоритма FastICA (подробнее см. п. 2.2-2.3) было получено следующее изображение (Рисунок 1). Дополнительно к нему прикрепляется таблица (Таблица 1), содержащая информацию об относительной близости кластеров друг к другу. Таблица была получена на основании применения метода косинусной близости (Схема 4). В основе этого метода лежит подсчёт косинуса угла между векторами, нормализованный произведением их длин.

$$\cos(\theta) = \frac{A * B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Схема 4. Формула подсчёта косинусного расстояния,
где A и B — векторы, соответствующие центрам кластеров, полученным при кластеризации.

Чем меньше значение в ячейке таблицы, тем ближе полученные кластеры друг к другу. Так, можно заметить, что ближе всего попарно оказались 4 и 5 кластеры (0,2362) и 3 и 4 кластеры (0,3229), то есть территории южнорусских говоров и севернорусских говоров оказываются ближе всего к среднерусским говорам. При этом примечательно, что 3 и 5 между собой находятся на большем расстоянии (0,6319). Кроме того, достаточно близкими оказываются украинские (№1) и белорусские (№2) говоры (0,4318), что противоречит (Дурново и др. 1915), предполагающему, что великорусские наречия ближе друг к другу, чем все прочие наречия русского языка. В свою очередь самыми отдалёнными оказываются 1 и 4, 1 и 5 кластеры (1,0 и 0,9639 соответственно), то есть украинские говоры и говоры севернорусские и среднерусские.

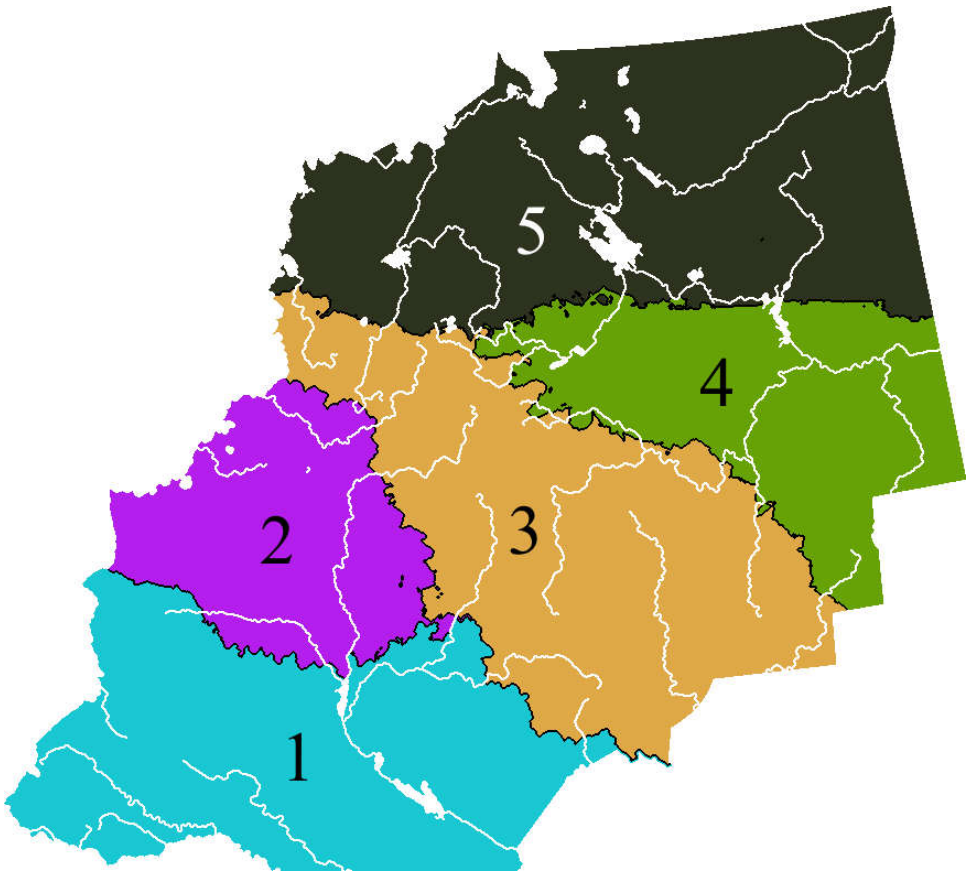

Рисунок 1. Кластеризация на 5 кластеров.

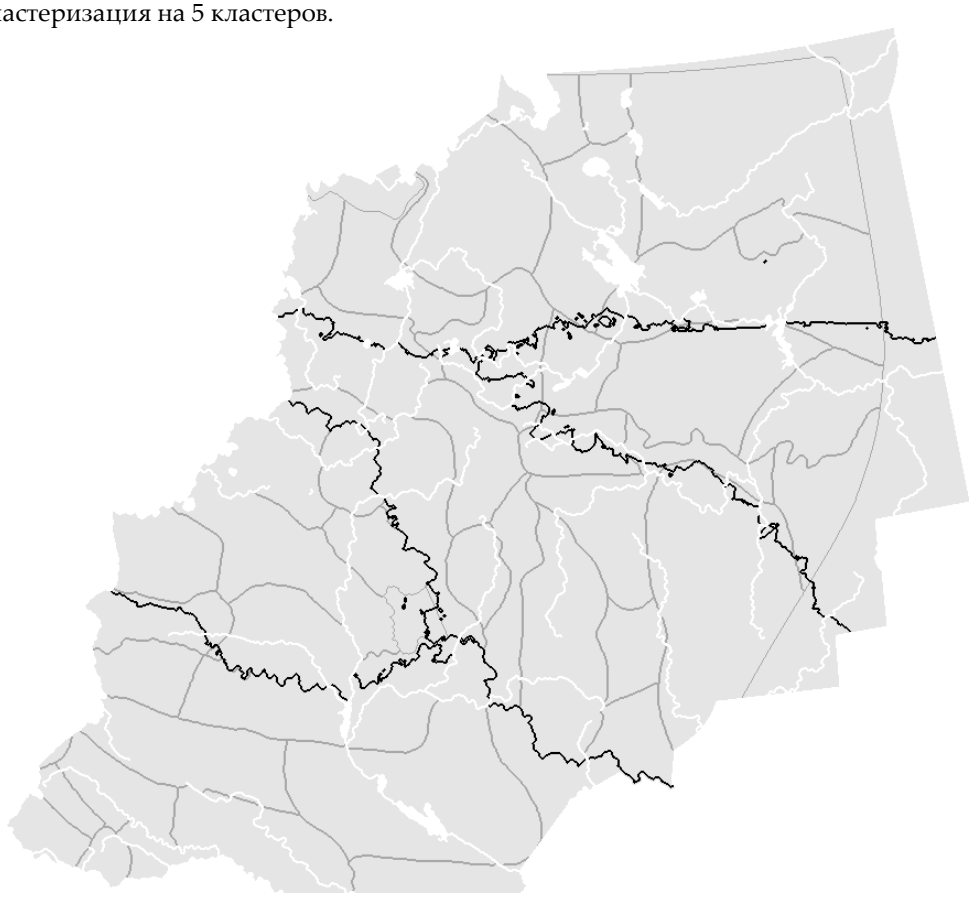

Рисунок 2. Наложение границ нашей классификации на 5 кластеров (черные границы) на границы классификаций русского (Захарова & Орлова, 1970), белорусского (Аванесаў и др. 1969) и украинского языка (Довгопол и др. 1977) (серые границы).

	1	2	3	4	5
1	0,0	0,4318	0,8788	1,0	0,9639
2	0,4318	0,0	0,5591	0,7778	0,7923
3	0,8788	0,5591	0,0	0,3229	0,4416
4	1,0	0,7778	0,3229	0,0	0,2362
5	0,9639	0,7923	0,4416	0,2362	0,0

Таблица 1. Попарная близость центров 5 кластеров, рассчитанная с помощью косинусного расстояния.

3.1.1. Сравнение кластеризации на 5 кластеров с существующими диалектными классификациями

Можно заметить, что на полученной карте политические границы почти совпадают с лингвистическими, однако трудно себе представить, что у жителей приграничных деревень, находящихся в десятке километров друг от друга, будут фиксироваться довольно разные диалектные черты. В особенности удивляет полное совпадение политической и лингвистической границы в районе северо-востока Украины и Курской, Белгородской областей России. Всё это может обуславливаться, в первую очередь, особенностями картографирования приграничных говоров в атласах ДАРЯ, ДАБМ, АУМ[2] и неединообразным отображением результатов[3]. Кроме того, на момент составления карт ВСИ третий том АУМ, в котором показаны говоры большей части Слобожанщины, Донетчины и Новороссии, ещё не был издан. В то же время, примечательно, что северно-полесские говоры, несмотря на перечисленные проблемы исходных данных, обнаруживают большую близость к говорам украинского языка (№1), а не к говорам белорусского языка (№2)[4]. Это согласуется с результатами работ (Аванесаў и др. 1969), (Довгопол и др. 1977).

Также важно отметить, что говор деревень на территории Брянского угла[5], относимый в классификации (Захарова, Орлова, 1970) к белорусским говорам, на нашей карте частично объединен с белорусским (№2) и с южнорусским кластерами (№3). Это может объясняться тем, что на данной территории наблюдается пересечение восточно-белорусских и западных южнорусских изоглосс (см. п. 3.2.3).

На территории русских говоров выделяются три крупных кластера, примерно соответствующих северному и южному наречиям и среднерусским говорам у (Захарова, Орлова, 1970). Опишем принципиальные различия (подробнее о диалектных признаках, участвующих в выделении этих кластеров, см. п. 3.1.2–3.1.3).

1. Традиционно выделяемое южнорусское наречие пополняется говорами, относимыми к среднерусским акающим. Так, Восточные среднерусские акающие говоры Отдела А в нашей кластеризации относятся большей частью к южнорусскому кластеру (№3). На западе границы южнорусского кластера также проходят значительно севернее, захватывая территории почти всей Псковской и юга Селиге-

[2] Например, на некоторых картах присутствуют доп. пометки об отсутствии точных картографированных данных (ВСИ-2, Карта 13, Признак 4), (ВСИ-3, Карта 11, Признак 3) и др.

[3] Например (ВСИ-3, Карта 2.1) «Наличие/отсутствие протетического гласного в последовательностях *#rъt, *#lъt», на которой распределение рефлексов происходит непосредственно по политическим границам.

[4] В п. 3.2.1. при описании кластеризации на 25 единиц, это будет рассмотрено более подробно.

[5] Западная часть Брянской области, граничащая с востоком Беларуси и севером Украины. Включает в себя несколько крупных населённых пунктов: Новозыбков, Стародуб, Клинцы, Унеча. Граница проходит на н./п. Почеп.

ро-торжковской групп говоров, традиционно относимых к Западным среднерусским акающим говорам.

2. Севернорусские говоры (№5) тоже расширились на западе относительно северного наречия в классификации (Захарова, Орлова, 1970). Это произошло благодаря присоединению территорий Западных среднерусских окающих говоров: Гдовской группы говоров и Новгородских говоров. Говоры Чухломского острова также стали частью севернорусского кластера (№5).
3. Оставшаяся территория объединяется в 4-й кластер, выделившийся последним от основной части севернорусского наречия, соответствует Восточным среднерусским говорам: в нашей кластеризации к этому кластеру были отнесены Восточные окающие (Владимирско-Поволжская группа) и, частично, Восточные акающие говоры: Отдел Б и Отдел В.

3.1.2. Определение наиболее значимых диалектных черт при выделении кластера

Далее для каждого кластера были определены 5 диалектных признаков (из 446, см. п. 2.1), внесших наибольший вклад в его выделение. В данной работе предлагается описание кластера №3; оценка проводилась с помощью метрики FM (см. п. 2.4). Результаты представлены в таблице (Таблица 2): в первом столбце содержится информация о номере диалектного признака в формате *том:карта:признак*. Во втором столбце дается название карты из сборника ВСИ, в третьем — признак на этой карте. Так, например, наибольший вклад в выделение третьего кластера внес второй признак «зелень, зеленя» на 18-й карте «Название (озимых) всходов зерновых культур» из второго тома ВСИ (ВСИ-2 1998) со значением FM индекса 0,8688. Он является наиболее значимым, так как почти полностью совпадает с выделившимся кластером, за исключением небольших спорадически распространённых по всему северу России и по северо-востоку Украины ареалов, где также встречается *зелень*, *зеленя*.

Примечательно, что в выделении этого кластера оказались значимыми и фонетические, и морфологические (именные и глагольные), и лексические признаки. При сравнении этих признаков с теми, которые выделяют для южного наречия (Захарова & Орлова 1970), наблюдается ряд совпадений. В частности, и в нашей работе, и в работе Захарова & Орлова 1970, ареал выделяется благодаря флексии /т'/ и недиссимилятивному аканью. Однако другие признаки, выделившиеся у нас, в работе Захарова & Орлова 1970 не встречаются. Например, признак «зелень, зеленя», лучше всего описывающий наш кластер, соответствует территориям южного наречия и западных среднерусских говоров, в этой работе при описании южнорусского наречия закономерно не упоминается.

3.1.3. Визуализация близости к кластеру №3

Предложенный в п. 3.1 подход к диалектной классификации восточнославянских языков предполагает дискретность языкового континуума, выделение ареалов как пучков изоглосс. Кроме этого, в нашей работе мы хотим предложить континуальный анализ ВСЯ, для которого была использована метрика Fowlkes–Mallows index (см. п. 2.4). Для рассматриваемого кластера (№3) рассчитывается метрика FM с каждым из 446 признаков. Таким образом получается массив из 446 значений в диапазоне [0, 1]. Затем каждый столбец исходной матрицы 406782*446 умножается на соответствующее значение массива, строки матрицы суммируются. В результате мы получаем массив длиной 406782,

каждое значение которого представляет собой меру близости, демонстрирующую, насколько в данной точке представлены признаки, хорошо приближающие рассматриваемый кластер. Наконец, генерируется изображение (Рисунок 3), цвета пикселей которого хорошо визуально различимы как «холодные» и «горячие». Чем «горячее» цвет пикселя, тем больше в нем представлено тех признаков, у которых есть пересечение с кластером.

№	Карта	Признак	FM index
2:18.2:2[6]	Название (озимых) всходов зерновых культур	*зелень*, *зеленя* и под.	0,8688
2:7.1:3	Флексия формы 3 лица ед. числа настоящего времени глаголов I спряжения	Флексия /т'/ (рус. диал. *он м'ёл'ет'*, *н'ес'ёт'*)	0,8139
3:1.1:4	Отношение к признакам вокальность / консонантность	Неразличение предударных *а* и *о* по модели недиссимилятивного аканья	0,7504
3:6:11	Флексия существительных Тв. п. мн. ч.	Изоглосса расширенного <м'и>[7]	0,7280
1:5:3	Морфонологическая характеристика консонантного исхода основы в парадигме наст. вр. глаголов I спр. типа русск. лит. *печь*, *сечь*, *течь*; *берёчь*, *стерёчь*, *стричь*	<К ~ К'> = <1 л. ед., 3 л. мн.> ~ <2, 3 л. ед., 1, 2 л. мн.> (например, *пеку́*, *пеку́т* ~ *пек'о́ш*...)	0,7234

Таблица 2. Наиболее значимые для кластера №3 диалектные признаки (оценка метрикой FM).

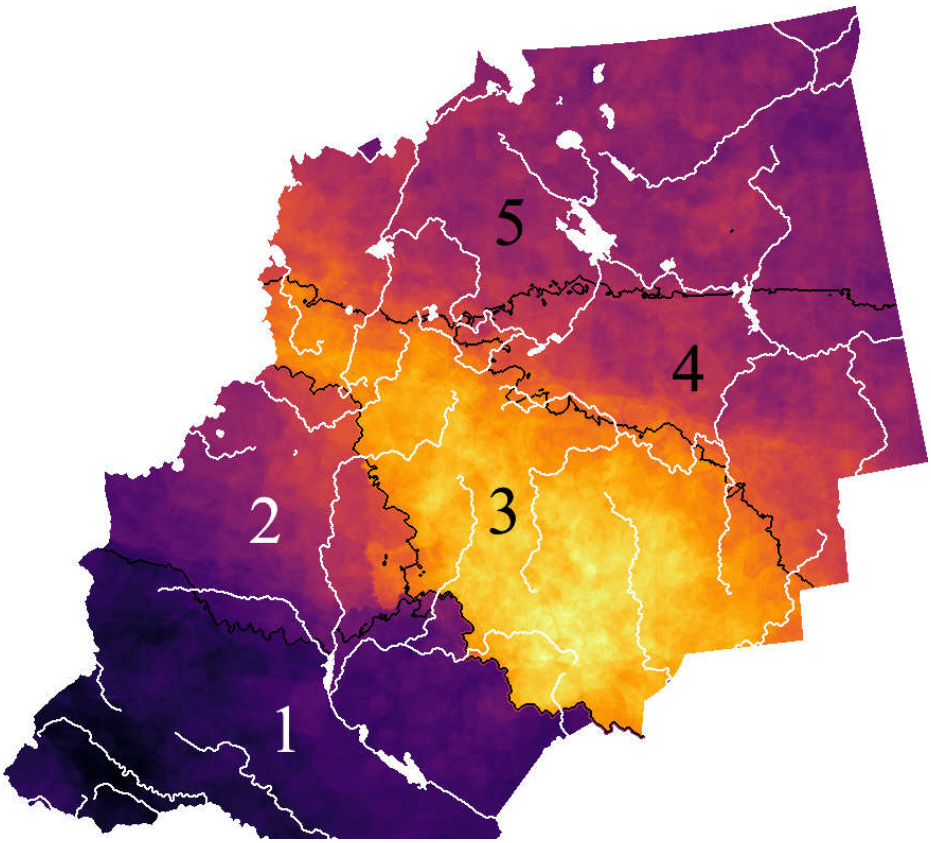

Рисунок 3. Близость к центру кластера №3 при кластеризации на 5.

[6] Изображение можно построить на электронном ресурсе Банка диалектологических карт 4-х томов ВСИ (Манусов и др. 2024).

[7] Окончание <м'и> не только наблюдается в словах *дверь*, *дочь*, *дети*, *лошадь*, *люди*, но и распространяется на другие слова.

Можно заметить, что ближе всего к исследуемому третьему кластеру оказываются территории среднерусских акающих говоров, восточная часть Беларуси и территория Брянского угла. Украинские территории, даже непосредственно граничащие с Россией, окрашены в очень темный цвет, что говорит о том, что на них представлено крайне мало тех признаков, которые представлены в исследуемом кластере, а те, что представлены, обладают низким значением FM. Говоры западной Беларуси и севернорусские говоры почти равноудалены от кластера №3.

Остановимся более подробно на диалектных признаках, наилучшим образом отдаляющих 3 кластер от соседних 2-го и 4-го, и признаках, описывающих их единство. Результаты представлены в таблицах (Таблица 3, Таблица 4).

При рассмотрении таблицы 3 можно заметить, что есть признаки, очень хорошо отдаляющие 3-й кластер от 2-го (ВСИ-2, Карта 18.2, Признак 2, FM=0,8629) и наоборот — 2-й от 3-го (ВСИ-2, Карта 14, Признак 3, FM=0,9262). Оба этих признака лексические — название озимых всходов зерновых культур *зелень*, *зеленя* и название деревянной посуды, в которой растворяют тесто *дзяжа́*, *дзёжка*, *дёжка*. В то же время нет признаков, которые бы достаточно хорошо описывали одновременно оба кластера, лучший признак — недиссимилятивное аканье — имеет FM=1,2327 (ВСИ-2, Карта 1.1, Признак 4) при максимуме 2. Это объясняется тем, что при кластеризации на малое количество кластеров разница между кластерами достаточно существенна, а сходства между ними невелики.

№	**Карта**	**Признак**	**FM index**
признаки, которые лучше всего описывают 3-й и 2-й кластеры вместе			
3:1.1:4	Отношение к признакам вокальность/консонантность	Неразличение предударных *а* и *о* по модели недиссимилятивного аканья	1,2327
4:5:2	Конечные /т/, /т'/ и ∅ в глагольной форме 3 л. мн. наст.	/т'/ (*несу́т'*)	1,1484
признаки, которые отличают 3-й кластер от 2-го			
2:18.2:2	Название (озимых) всходов зерновых культур	*зелень*, *зеленя* и под.	0,8629
3:6:11	Флексия творительного падежа множественного числа существительных	Изоглосса расширенного <м'и>	0,7280
признаки, которые отличают 2-й кластер от 3-го			
2:14:3	Названия деревянной посуды, в которой растворяют тесто	*дзяжа́*, *дзёжка*, *дёжка*	0,9262
3:2.1:2	Наличие/отсутствие протетического гласного в последовательностях *#rъt, *#lъt	*ры*, *лы*, *ли*	0,9262

Таблица 3. Диалектные признаки, описывающие 3-й кластер в сравнении с 2-м (оценка метрикой FM).

При сравнении 3-го и 4-го кластеров можно заметить, что 3-й кластер также значительно отличается от 4-го, причем признак, который лучше всего отличает 3-й кластер от 4-го тот же, что и при отличии 3-го от 2-го — название озимых всходов зерновых культур *зелень*, *зеленя*. В свою очередь отличия 4-го кластера от 3-го менее существенны (наибольшее значение метрики FM=0,6987 для формы Тв.п. числительного *два* в зоне разреженного распространения *-ми*), что может быть свидетельством того, что 4-й кластер выделился скорее не по набору своих уникальных черт, а по набору уникальных черт у соседних 3-го и 5-го кластера и по их отсутствию у 4-го. Лучше всего описывают 3-й и 4-й кластер вместе форма Тв.п. числительного *два* с окончанием *-ми* (ВСИ-1, Карта 8, При-

знак 5), для которого FM=1,2190 и название укладок снопов из рефлексов слав. **krъstъ* (ВСИ-4, Карта 8.2, Признак 2), для которого FM=1,1562.

№	Карта	Признак	FM index
признаки, которые лучше всего описывают 3-й и 4-й кластеры вместе			
1:8:5	Форма творительного падежа числительного *два*	Окончание *-ми*	1,2190
4:8.2:2	Название укладок снопов (в поле)	Рефлексы слав. **krъstъ*: рус. диал. *к(х)рестец*, *к(х)рест*, *к[х]рестик* и др.; укр. диал. *хрест*, *хрестик*, *хрест'*, *христец'*, *хрест'іўка* и др.; бел. диал. редк. *(у) кры'жык'и*	1,1562
признаки, которые отличают 3-й кластер от 4-го			
2:18.2:2	Название (озимых) всходов зерновых культур	*зелень*, *зеленя* и под.	0,8001
2:7.1:3	Флексия формы 3 лица ед. числа настоящего времени глаголов I спряжения	Флексия /т'/ (рус. диал. *он м'ёл'ет'*, *н'ес'ёт'*)	0,7829
признаки, которые отличают 4-й кластер от 3-го			
1:8:6	Форма творительного падежа числительного *два*	Зона разреженного распространения *-ми*	0.6987
2:4.1:5	Утрата [j] и стяжение гласных в глагольных формах	Формы глаголов типа рус. *бол[é]ш*, *бол[é]т*, *бол[é]м* и т. п.	0.5956

Таблица 4. Диалектные признаки, описывающие 3-й кластер в сравнении с 4-м (оценка метрикой FM).

3.2. Кластеризация на 25 кластеров

При кластеризации на 25 кластеров выделяется 6 кластеров на территории Украины, 5 — на территории Беларуси и 14 в России. К кластеризации прикрепляется таблица (подробнее см. п. 3.1). В данной таблице значения будут ближе к нулю, чем при разделении на 5 кластеров, то есть соседние кластеры будут ближе друг к другу. В таблице наиболее примечательным является 5 кластер. Он показывает наименьшую близость к другим кластерам на территории Украины и одновременно с этим оказывается ближе всех украинских кластеров к русским кластерам, особенно разница заметна при сравнении других украинских кластеров с севернорусскими говорами: между 5 и 24 кластером FM=0,81, между 1 и 24 кластером FM=0,95.

3.2.1. Сравнение кластеризации на 25 кластеров с существующими диалектными классификациями

При сравнении карты, построенной на 25 кластеров с диалектным членением в работе (Захарова & Орлова 1970), можно заметить следующие параллели:

1. Границы севернорусского и южнорусского наречия, среднерусских говоров, почти совпадают.
2. Многие выделяемые ареалы находятся в тех же границах, что и в работе Захарова & Орлова 1970, это, в частности, Псковская группа говоров (кластер №20), Костромская группа (№24), Владимирско-поволжская группа (№18), Рязанская группа (№16), Западная группа говоров (№12) за исключением Брянского угла, относимого в работе (Захарова, Орлова 1970) к белорусским говорам.

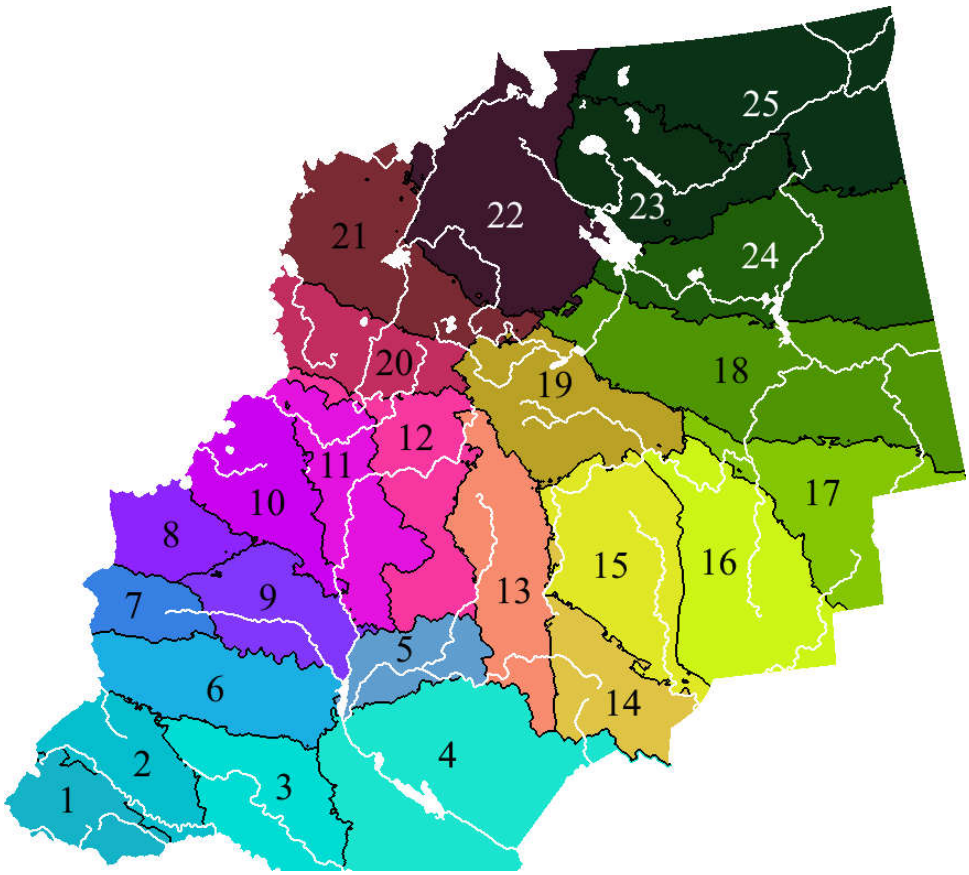

Рисунок 4. Кластеризация на 25 кластеров.

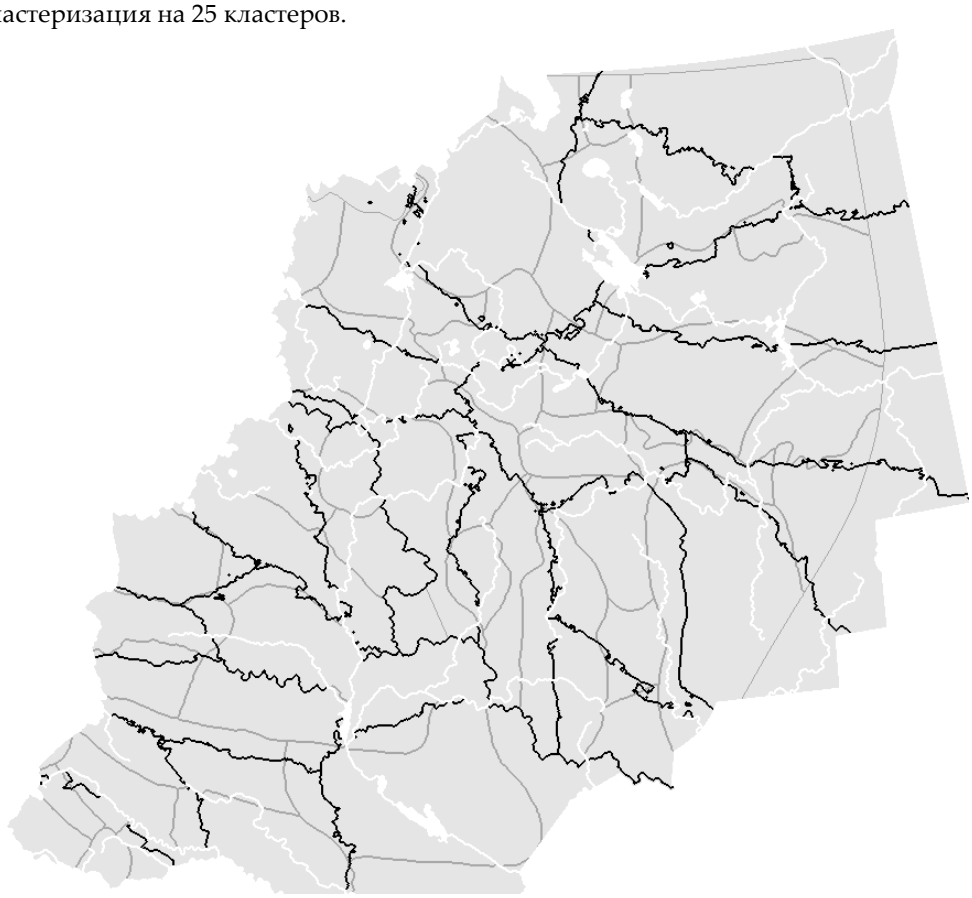

Рисунок 5. Наложение границ нашей классификации на 25 кластеров (черные границы) на границы классификаций русского (Захарова & Орлова, 1970), белорусского (Аванесаў и др. 1969) и украинского языка (Довгопол и др. 1977) (серые границы).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	0,00	0,16	0,30	0,35	0,49	0,36	0,42	0,52	0,54	0,64	0,68	0,82	0,88	0,88	0,97	0,99	0,98	0,94	0,95	0,90	0,95	0,92	0,91	0,95	0,92
2	0,16	0,00	0,19	0,34	0,50	0,29	0,41	0,49	0,53	0,64	0,70	0,84	0,88	0,90	0,98	1,00	1,00	0,96	0,96	0,92	0,97	0,94	0,92	0,94	0,94
3	0,30	0,19	0,00	0,19	0,36	0,14	0,30	0,47	0,43	0,62	0,65	0,79	0,84	0,86	0,92	0,94	0,95	0,93	0,93	0,91	0,96	0,92	0,91	0,93	0,92
4	0,35	0,34	0,19	0,00	0,19	0,20	0,32	0,51	0,43	0,58	0,58	0,70	0,73	0,75	0,84	0,88	0,93	0,90	0,89	0,88	0,93	0,89	0,88	0,88	0,88
5	0,49	0,50	0,36	0,19	0,00	0,23	0,30	0,42	0,31	0,41	0,41	0,56	0,64	0,67	0,73	0,75	0,81	0,81	0,75	0,76	0,83	0,82	0,82	0,81	0,83
6	0,36	0,29	0,14	0,20	0,23	0,00	0,16	0,39	0,32	0,52	0,56	0,72	0,80	0,84	0,90	0,93	0,95	0,92	0,92	0,90	0,96	0,91	0,88	0,90	0,88
7	0,42	0,41	0,30	0,32	0,30	0,16	0,00	0,30	0,25	0,44	0,51	0,70	0,80	0,82	0,88	0,90	0,92	0,89	0,89	0,88	0,94	0,90	0,84	0,87	0,84
8	0,52	0,49	0,47	0,51	0,42	0,39	0,30	0,00	0,18	0,18	0,38	0,60	0,70	0,77	0,79	0,83	0,85	0,85	0,77	0,73	0,84	0,84	0,85	0,86	0,86
9	0,54	0,53	0,43	0,43	0,31	0,32	0,25	0,18	0,00	0,26	0,33	0,56	0,69	0,74	0,77	0,79	0,81	0,82	0,75	0,74	0,84	0,83	0,81	0,81	0,82
10	0,64	0,64	0,62	0,58	0,41	0,52	0,44	0,18	0,26	0,00	0,14	0,40	0,55	0,64	0,66	0,71	0,74	0,73	0,62	0,60	0,72	0,75	0,78	0,77	0,79
11	0,68	0,70	0,65	0,58	0,41	0,56	0,51	0,38	0,33	0,14	0,00	0,23	0,42	0,53	0,57	0,61	0,68	0,69	0,51	0,51	0,64	0,68	0,73	0,74	0,76
12	0,82	0,84	0,79	0,70	0,56	0,72	0,70	0,60	0,56	0,40	0,23	0,00	0,16	0,36	0,44	0,47	0,54	0,58	0,39	0,29	0,48	0,56	0,63	0,63	0,64
13	0,88	0,88	0,84	0,73	0,64	0,80	0,80	0,70	0,69	0,55	0,42	0,16	0,00	0,18	0,29	0,36	0,48	0,53	0,31	0,35	0,48	0,56	0,63	0,60	0,64
14	0,88	0,90	0,86	0,75	0,67	0,84	0,82	0,77	0,74	0,64	0,53	0,36	0,18	0,00	0,12	0,26	0,44	0,51	0,33	0,39	0,43	0,52	0,58	0,56	0,60
15	0,97	0,98	0,92	0,84	0,73	0,90	0,88	0,79	0,77	0,66	0,57	0,44	0,29	0,12	0,00	0,11	0,29	0,42	0,20	0,41	0,43	0,52	0,56	0,52	0,57
16	0,99	1,00	0,94	0,88	0,75	0,93	0,90	0,83	0,79	0,71	0,61	0,47	0,36	0,26	0,11	0,00	0,19	0,37	0,20	0,44	0,47	0,55	0,54	0,47	0,56
17	0,98	1,00	0,95	0,93	0,81	0,95	0,92	0,85	0,81	0,74	0,68	0,54	0,48	0,44	0,29	0,19	0,00	0,14	0,18	0,44	0,39	0,43	0,39	0,31	0,39
18	0,94	0,96	0,93	0,90	0,81	0,92	0,89	0,85	0,82	0,73	0,69	0,58	0,53	0,51	0,42	0,37	0,14	0,00	0,22	0,46	0,34	0,30	0,26	0,16	0,31
19	0,95	0,96	0,93	0,89	0,75	0,92	0,89	0,77	0,75	0,62	0,51	0,39	0,31	0,33	0,20	0,20	0,18	0,22	0,00	0,31	0,3	0,37	0,42	0,35	0,48
20	0,90	0,92	0,91	0,88	0,76	0,90	0,88	0,73	0,74	0,60	0,51	0,29	0,35	0,39	0,41	0,44	0,44	0,46	0,31	0,00	0,18	0,33	0,48	0,47	0,51
21	0,95	0,97	0,96	0,93	0,83	0,96	0,94	0,84	0,84	0,72	0,64	0,48	0,48	0,43	0,43	0,47	0,39	0,34	0,30	0,18	0,00	0,11	0,31	0,33	0,36
22	0,92	0,94	0,92	0,89	0,82	0,91	0,90	0,84	0,83	0,75	0,68	0,56	0,56	0,52	0,52	0,55	0,43	0,30	0,37	0,33	0,11	0,00	0,17	0,24	0,22
23	0,91	0,92	0,91	0,88	0,82	0,88	0,84	0,85	0,81	0,78	0,73	0,63	0,63	0,58	0,56	0,54	0,39	0,26	0,42	0,48	0,31	0,17	0,00	0,13	0,08
24	0,95	0,94	0,93	0,88	0,81	0,90	0,87	0,86	0,81	0,77	0,74	0,63	0,60	0,56	0,52	0,47	0,31	0,16	0,35	0,47	0,33	0,24	0,13	0,00	0,18
25	0,92	0,94	0,92	0,88	0,83	0,88	0,84	0,86	0,82	0,79	0,76	0,64	0,64	0,6	0,57	0,56	0,39	0,31	0,48	0,51	0,36	0,22	0,08	0,18	0,00

Таблица 5. Попарная близость центров 25 кластеров, рассчитанная с помощью косинусного расстояния.

3. Также есть кластеры, состоящие из нескольких ареалов на карте Захарова & Орлова 1970: объединены Гдовская группа и Новгородские говоры (№21); Восточные среднерусские акающие говоры (Отдел Б и В) (№17); Верхне-Днепровская, Верхне-Деснинская группы и Межзональные говоры типа А (№13), что частично объясняется меньшим количеством классификационных единиц в нашей работе по сравнению с тем, сколько насчитывается единиц в диалектных членениях русского, белорусского и украинского языков вместе. Примечательно, что при дополнительном построении на 50 кластеров[8] кластер №17 остается в тех же границах без дополнительных членений. Кластер №21 разбивается на 2, граница между ними проходит восточнее, чем на карте (Захарова, Орлова 1970), примерно по о. Ильмень. Кластер №13 разбивается на 2 кластера вместо традиционных трёх, при этом граница проходит к востоку от Оки.

[8] Не публикуем кластеризацию в данной работе, см. п. 2.2.

Наблюдается и ряд расхождений:

1. Самое заметное — в нашем диалектном членении, построенном при равном весе признаков из разных доменов языка, Чухломский остров не выделяется. Более того, при кластеризации на 50 кластеров он также не выделяется в самостоятельный кластер, а образует ареал, с севера ограниченный Чухломой, а с юга — Горьковским водохранилищем. Мы находим этому следующие объяснения. Во-первых, Чухломской остров на карте Захарова & Орлова 1970 в первую очередь выделяется по набору фонетических черт (в основном связанных с реализацией гласных в безударных слогах, что является важной изоглоссой, разграничивающей русские диалекты), в то время как в ВСИ из 18-ти фонетических карт только две посвящены гласным в безударной позиции[9], то есть ВСИ не представлен тот языковой материал, который выделяет Чухломской остров в отдельный ареал на основе карт ДАРЯ. Во-вторых, черты из других разделов языка, в первую очередь, из морфологии, которые объединяют эту территорию с среднерусскими Владимирскими говорами по материалам ДАРЯ, также мало представлены в ВСИ. К таким чертам относится, например, мягкость задненёбного согласного основы в Пред.п. ед.ч. прил. м. и ср. р. с безударными окончаниями (ДАРЯ, Вып. II, Карта 47). На материале ВСИ из 24 морфологических и морфонологических карт нет тех, на которых бы Чухломской остров выделялся в отдельный ареал или объединялся в ареал со среднерусскими говорами, что мы могли бы от него ожидать. На морфологических картах ВСИ Чухломской остров наоборот часто становится частью севернорусского наречия.
2. Ареалы на территории севернорусских говоров (в частности, северные межзональные говоры) не выделились. Вместо этого Ладого-Тихвинская группа говоров протянулась до Онежского озера на севере и до Рыбинского водохранилища на востоке (№22). Признак, который лучше всего отличает кластер №22 от кластера №23 — название орудия для ручного обмолота *прúвяз*, *прúвязь* (ВСИ-4, Карта 7.1, Признак 5). Кластеры №23 и №25 отличаются друг от друга минимально (Таблица 2), поэтому и признаки, которые различают их, делают это достаточно слабо. Так, 25-й кластер отличается от 23-го по названию валька для выколачивания белья — *палка* (ВСИ-2, Карта 18.1, Признак 22). При кластеризации на 50 кластеров от Ладого-Тихвинских говоров отделяется кластер, включающий в себя Онежские и Лачские говоры.
3. Территории вокруг Вологды, Череповца, Белозерска объединились в кластер №23, которого нет на карте 1965 г. Данный кластер выделяется, в первую очередь, по тому, как в нем реализовались рефлексы *'а и *ę (ВСИ-1, Карта 4, Признак 8): как *е* перед мягкими согласными (*p'ét'*, *s'éd'eš*), и как *а* перед твердыми (*p'átyj*, *s'ádu*). Кроме того, на выделение повлияли формы указательных местоимений. В большей части этого кластера а также частично на территории 22-го и 25-го кластеров для указания на ближний и отдаленный предмет используется местоимение *тот* или *э́тот* (ВСИ-3, Карта 7.2, Признак 13). При кластеризации на 50 кластеров выделились два кластера, располагающихся на территориях Белозерско-Бежецких говоров.

[9] Тем не менее, при создании кластеризации на 5 на основании исключительно фонетических признаков Чухломской остров выделяется и объединяется со среднерусскими говорами, так как отмечается на карте (ВСИ-4, Карта 2.1, Признак 2 — недиссимилятивное аканье). При кластеризации на 25 он не выделяется.

4. Также в центре карты образовался кластер №19, объединяющий части Селигеро-Торжковских говоров, Тверской подгруппы Владимирско-Поволжской группы и Отдела А Восточных среднерусских акающих говоров. Он в тех же границах сохраняется и при кластеризации на 50. Было обнаружено, что данный кластер имеет ряд специфических черт, отличающих его от соседних кластеров. Самая значимая из них, которая больше всего отличает 19 кластер от 15–18 кластеров — название бьющей части цепа — *бичу́к*, *би́чик* (ВСИ-4, Карта 7.3), в то время как на территории 15-го кластера она называется *тепи́нка*, *цепи́нка*, 16-го — *тепе́ц*, *тепо́к*, 17-го — *цепе́ц*, *цепи́льник*, *цепни́к*.

При сравнении построенных на территории Украины кластеров с ареалами, выделяемыми качественными методами на картах (Довгопол 1977) также наблюдается ряд сходств и отличий:

1. На территории Украины отчетливо выделяются три основных наречия: Северное, Юго-Восточное и Юго-Западное.
2. Полесские говоры на территории Северного наречия делятся на Восточные (№5) и Западные вместе со Средними (№6). При кластеризации на 50 кластеров кластер №6 делится на Западно- и Среднеполесские говоры.
3. В Юго-Западном наречии выделяются в один кластер Подольские говоры и Восточная часть Волынских говоров (№3), в другой кластер вошли Поднестровские говоры и Западная часть Волынских говоров (№2). В третью группу по правую руку от р. Днестр вошли Закарпатские, Бойковские, Гуцульские, частично Покутско-Буковинские говоры (№1). При большем количестве кластеров диалектное членение украинских говоров становится значительно более подробным: начинают выделяться в отдельный кластер Закарпатские и Бойковские говоры, в другой — Покутско-Буковинские и Гуцульские говоров.
4. Юго-восточное наречие при кластеризации на 25 кластеров внутри себя не делится на кластеры. При более подробной кластеризации выделяются Слобожанские говоры, а также две территории с границей по линии Полтава-Черкасы.

В Беларуси при сравнении с (Аванесаў и др. 1969) выделяются:

1. Северные Полесские говоры (№7), расположенные между соседними белорусскими с севера (№8) и украинскими говорами с юга (№6). При этом при оценке близости кластер №7 оказывается значительно ближе именно к украинскому шестому кластеру (косинусное расстояние 0,16), а не к белорусскому (косинусное расстояние 0,30). Лучше всего 6-й и 7-й кластер описывается использованием Род.п. в позиции объекта: укр. *уз'аў ножа́*, бел. *знайшоў гры́ба* (ВСИ-2, Карта 11, Признак 2), FM=0.9281. При этом признак, лучше всего описывающий 7-й и 8-й кластеры — указательное местоимение ближнего дейксиса м.р. ед.ч. *ге́той* — имеет значительно более низкое значение FM=0.7438. Это согласуется с работой (Аванесаў и др. 1969: 1-9). В ней выделяется большое количество уникальных, не встречающихся на других территориях Беларуси, диалектных черт Северного Полесья, в первую очередь на основе морфологических и лексических черт: например, указательное местоимение Твор.п. *за тэ́ю*, прилагательное Мест.п. ед.ч. м.р. *об маладо́му* — как в литературном украинском (при этом в литературном белорусском — *аб малады́м*).
2. Выделяются кластеры №8 и №9. Вместе они соответствуют Северо-западной и Мозырьской группам говоров, относящимся к Югу-западному диалекту.

3. На северо-востоке выделяются кластеры №10 и №11, первый из которых соответствует Полоцкой, а второй Витебско-Могилевской группе говоров Северо-восточного диалекта. Кластер №11 формируется, частично, за счёт трансграничных диалектных черт, которые помимо восточной Беларуси также представлены на юго-западе России. Это, например, указательные местоимения м.р. ед.ч. дальнего дейксиса *тэй*, *тый* (ВСИ-3, Карта 7.1, Признак 3), протетический губной согласный (*в*, *w*) или его отсутствие перед инициальным гласным *о* (ВСИ-2, Карта 1, Признак 2), протетический гласный в последовательностях *#rъt, *#lъt — *иржать*, *ильна* (ВСИ-3, Карта 2, Признак 19). В свою очередь кластер №10 лучше всего описывается диалектными чертами, представленными непосредственно внутри Беларуси. Это, например, указательные местоимения м.р. ед.ч. дальнего дейксиса *[ӱэ́]ный*, *[ӱэ́]ндой* (ВСИ-3, Карта 7.1, Признак 4), указательные местоимения м.р. ед.ч. ближнего дейксиса *гэ́той* (ВСИ-3, Карта 7.2, Признак 4), название лемеха *нарог* (ВСИ-3, Карта 13.1, Признак 4).

3.2.2. Определение наиболее значимых диалектных черт при выделении кластера

Для более подробного описания в этой работе был выбран ареал №12, объединяющий южную часть Псковской и юго-западную часть Тверской областей (по линии Себеж—Нелидово), западную часть Смоленской области (линия Холм-Жирковский—Сафоново—Десногорск), западную часть Брянской области, примерно ограниченную Десной. Наибольшее значение FM в данном списке не превышает 62%, в сравнении с п. 3.1.2, в котором для 5 кластеров FM достигает 87%. Это обуславливается тем, что при увеличении количества кластеров качество генерации постепенно снижается. Примечательно, что тремя самыми значимыми признаками, участвующими в выделении данного кластера, являются признаки, связанные с реализацией гласного в предударном слоге, а именно — с диссимилятивным аканьем. Так как все признаки имеют равный вес в генерации, значимым оказался признак единичной встречаемости форм *кони поить*, *коровы доить*. Самой значимой лексической чертой является название лемеха как рефлекса слав. **lemešь* (*лемех*, *лемеш*).

3.2.3. Визуализация близости к кластеру №12

О том, как было получено изображение (Рисунок 6) и для чего оно предназначается, написано в п. 3.1.3. Заметно, что ближе всего к центру 12-го кластера оказываются точки соседнего 13-го кластера (особенно, его западная часть). 13-й кластер занимает территории запада Белгородской области, западной половины Курской области, западной части Орловской области (к югу от Орла граница ареала совпадает с течением Оки, далее по линии Орёл—Болхов), юго-западной половины Калужской области (по линии Козельск—Юхнов), в Смоленской тянется тонкой полосой, включающей Вязьму, до самой границы с Тверской областью, восточной части Брянской области.

В таблице 6 содержится информация о том, какие диалектные признаки лучше всего приближают 12-й и 13-й кластеры, а какие — лучше всего их отдаляют. Заметно, что лучше всего кластеры отличаются по лексическим признакам (названию ручки цепа, кукушки, валька для выколачивания белья). Это объясняется в первую очередь тем, что лексические ареалы часто значительно меньше, чем фонетические, морфологические и синтаксические и поэтому лучше описывают кластеры меньшего размера. Нет признаков, которые бы отличали кластеры более чем на 50%, так как чем меньше становятся кластеры, тем хуже они объясняются одним отдельно взятым признаком.

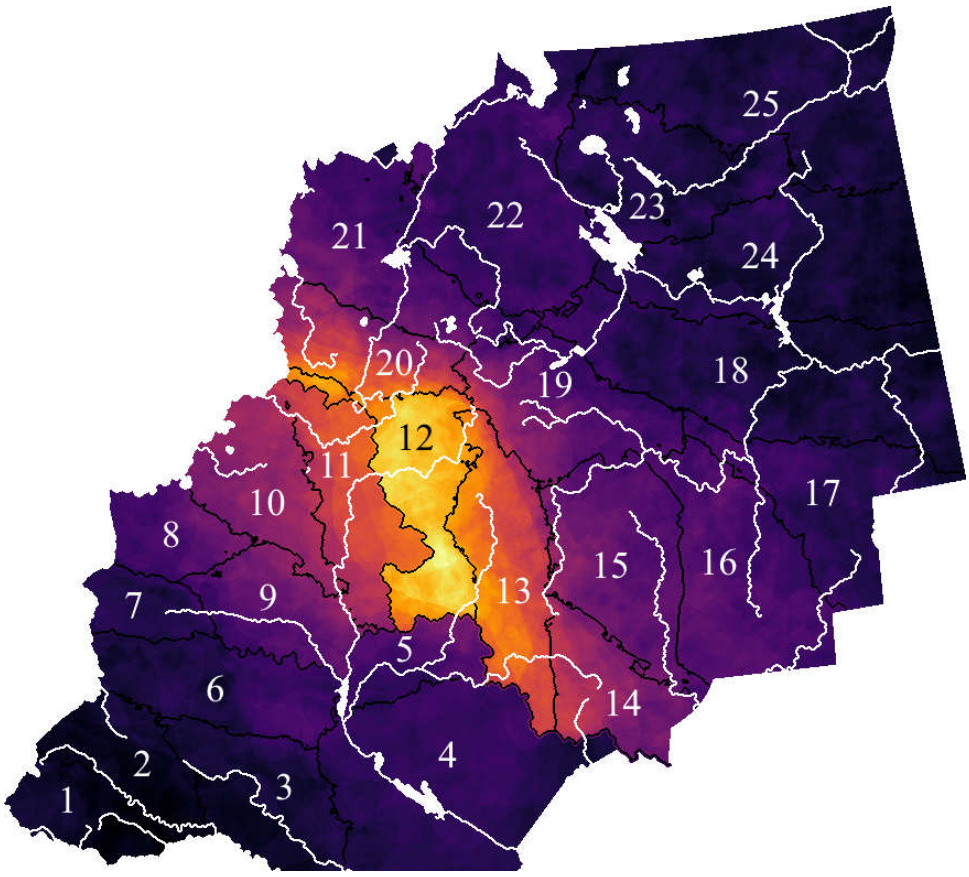

Рисунок 6. Близость к центру кластера №12 при кластеризации на 25.

№	Карта	Признак	FM index
4:2.1:7	Непередние/задние гласные в предударном слоге	При диссимилятивном аканье изменение *ъ/ы* в *у* в соседстве с губными согласными: *пъдво́з*, но *пум'аў*, *пубра́ў*, *кумпа́н'ийа*	0,6171
4:6.6:3	Реализация тематического гласного в 1 л. мн. наст. ([о] или [не-о])	Зона диссимилятивного аканья на русской территории	0,5737
3:1.1:5	Отношение к признакам вокальность/консонантность	Неразличение предударных *а* и *о* по модели диссимилятивного аканья: *во́ды* > *вадо́й*, *въда́*; *са́м* > *само́й*, *съма́*)	0,5718
3:12:7	Форма Вин.п. мн.ч. одушевленных существительных	Формы Вин.-Им. п. отмечены в единичных пунктах: рус. *кони поить*, *коровы доить*	0,5717
3:13.1:2	Названия лемеха	Преимущественное распространение рефлексов слав. **lemešь*: *лемех*, *лемеш*	0,5540

Таблица 6. Наиболее значимые для кластера №12 диалектные признаки (оценка метрикой FM).

Кроме того, по рисунку 4 видно, что 12 кластер также оказывается близок к 11-му, восточнобелорусскому, кластеру. В обоих кластерах широко представлены следующие диалектные явления: употребление объектно-целевых конструкций типа *в грибы*, *ў грыбы* (ВСИ-1, Карта 11, Признак 5), указательные местоимения *тэй*, *тый* (ВСИ-3, Карта 7, Признак 3), диссимилятивное аканье в предударном слоге (ВСИ-4, Карта 2.1, Признак 3). Близость 12-го кластера к восточнобелорусским говорам отмечается во многих, в том числе ранних работах, посвященных диалектному членению восточнославянского ареала (Дурново и др. 1915); это же подтверждается в работе (Пшеничнова 1996), в которой западные говоры на границе с Белоруссией отделяются от прочих русских говоров уже на

первом уровне деления. Однако в силу нехватки в нашей работе карт, значимых для диалектного членения восточнославянских языков, 12 кластер — часть русского ареала, близкий не только к 13-му, но и к 14-му, южнорусскому, и 20, западнорусскому, кластерам.

Наиболее удалёнными оказываются территории юго-запада Украины и севера России.

№	Карта	Признак	FM index
признаки, которые *лучше* всего описывают 12-й и 13-й кластеры вместе			
4:6.6:3	Реализация тематического гласного в 1 л. мн. наст. ([о] или [не-о])	Зона диссимилятивного аканья на русской территории	1,1926
3:12:7	Форма Вин.п. мн.ч. одушевленных существительных	Формы Вин.-Им. п. отмечены в единичных пунктах: рус. *кони поить, коровы доить*	1,1263
признаки, которые отличают 12-й кластер от 13-го			
4:7.2:3	название ручки цепа	*цепильно́, цепилно́*	0,4919
2:15:12	название кукушки	*зезю́лька, зозу́лька*	0,4213
признаки, которые отличают 13-й кластер от 12-го			
2:18.1:3	название валька для выколачивания белья	*пра́льник*	0,4755
2:5:5	Морфонологическая характеристика консонантного исхода основы в парадигме настоящего времени глаголов II спряжения типа рус. лит. *любить*	<P'> = <1, 2, 3 л. ед., 1, 2, 3 л. мн.>: *л'уб'у́ ~ л'у́б'иш..., л'у́б'ат'*	0,4460

Таблица 7. Диалектные признаки, описывающие 12-й кластер в сравнении с 13-м (оценка метрикой FM).

3.3. Диалектно-континуальный подход к классификации ВСЯ

Нам кажется важным провести параллель между нашим исследованием и работой (Пшеничнова 1996). В диалектном членении Н. Н. Пшеничновой делается попытка одновременно отразить *непрерывность* и *дискретность* языкового континуума для русских говоров первичного формирования. Это становится возможным благодаря выделению классификационных единиц разных уровней: единицы первого и второго уровня, находящиеся вверху иерархии, показывают черты дискретности, единицы более низких уровней, третьего и четвертого — континуальности.

Наши классификации в п. 3.1.1 и п 3.2.1 являются дискретными, так как создают ареалы из пучков изоглосс, однако использование цветов для отображения близости кластеров друг к другу позволяет получить больше информации о взаимосвязях между кластерами. Карты, которые мы предлагаем для оценки близости всех точек на карте к центру конкретного выбранного кластера (п. 3.1.3 и п. 3.2.3) являются хорошим примером континуального подхода, при котором «угасание» показывает сокращение совпадающих диалектных признаков. В этом разделе мы предлагаем вариант дискретно-континуального описания восточнославянского ареала.

Для его построения использовалось Евклидово расстояние[10] (Схема 5) — кратчайшее расстояние между двумя векторами в *n*-мерном пространстве. Оно рассчитывалось между каждой точкой кластера и его центром. Затем значения для точек каждого кластера были нормализованы, то есть приведены к диапазону [0, 1]. Полученные значения

[10] От косинусного расстояния в этом разделе пришлось отказаться, так как оно не чувствительно к длине векторов, существенной при построении дискретно-континуальных карт.

визуализированы на картах (Рисунок 7, Рисунок 8). Наиболее близкие к центру кластера точки имеют более светлый цвет, самые отдаленные точки показаны более темными оттенками. Следовательно, точки на границе нескольких кластеров являются темными — переходными — зонами, в которых меньше всего представлены черты центра кластера и представлены черты соседних диалектных единиц.

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

Схема 5. Формула подсчёта Евклидова расстояния,
где p_n — вектор кластера p в n-мерном пространстве, q_n — вектор кластера q в n-мерном пространстве.

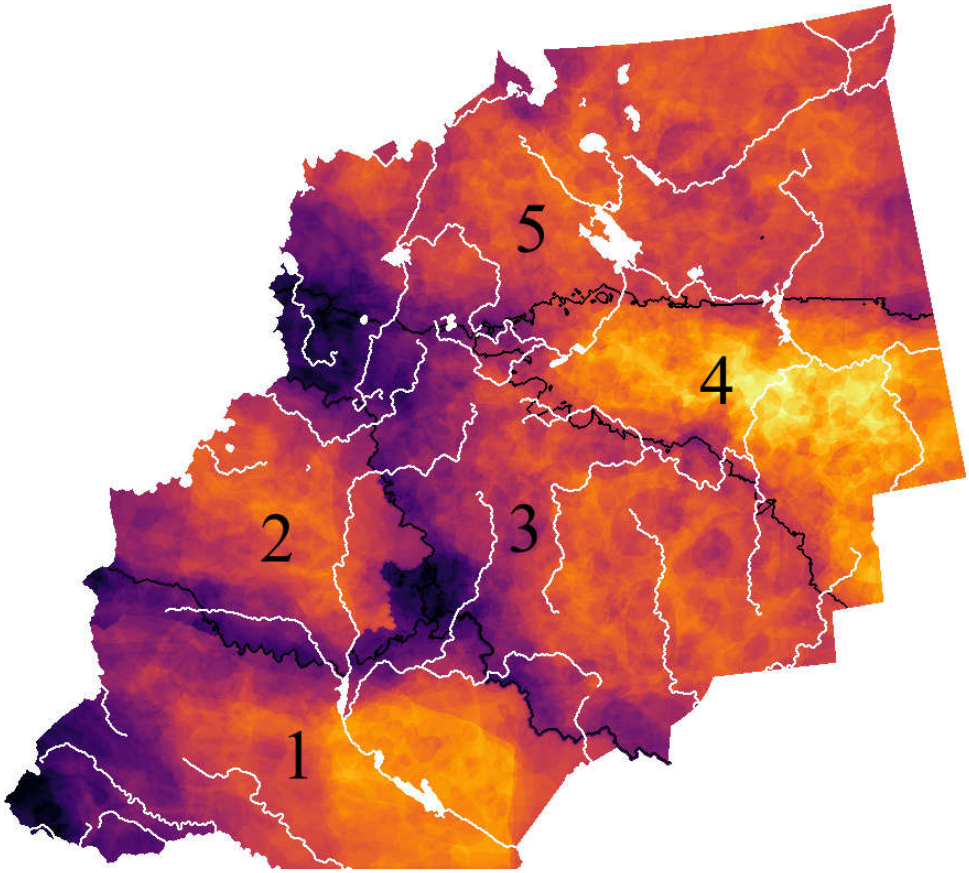

Рисунок 7. Дискретно-континуальный подход к классификации ВСЯ на 5 диалектных единиц.

Поскольку нормализация проводится на значениях всех точек сразу, существует сильная тенденция к тому, чтобы при кластеризации на N признаков выше других были значения в кластере, выделившимся последним, в силу того что он включает в себя точки, наиболее похожие друг на друга и вместе с тем отличающиеся от центров своих кластеров при кластеризации на N-1 кластеров. Выделяясь, такой кластер увеличивает значения и в кластерах, из которых он выделился. Важно отметить, что область внутри границ третьего кластера визуально отличается от той же области на изображении с визуализацией расстояний до центра только третьего кластера (Рисунок 3), за что также ответственны иная метрика и единая нормализация по всем значениям. Если бы нормализация проводилась независимо по каждому кластеру, контраст между разными кластерами был бы утрачен.

Ярче других при кластеризации на 5 (Рисунок 7) горит 4-й кластер, так как он выделился последним и, следовательно, на данный момент является самым хорошо при-

ближенным своими признаками. Самые темные территории свидетельствуют о том, что для них наблюдается наименьшее количество общих черт с центром кластера, а значит, при генерации на большее число кластеров они первыми распадутся и образуют новые диалектные единицы. Так, заметно, что самыми темными являются области возле Пскова и Брянска, а также территория на западе Украины. И действительно, 6-й выделившийся кластер — территория Псковских, Гдовских, Новгородских и Ладого-Тихвинских говоров, 7-й кластер соответствует Юго-западным русским говорам, 8-й делит Украину на запад и восток, а 9-й выделяет из Западных украинских говоров Юго-западные. Таким образом, можно сказать, что наш метод обладает предсказательной силой, позволяя не только предположить более подробное диалектное членение на основании более крупных диалектных единиц, но и определить последовательность возникновения дроблений внутри этих единиц.

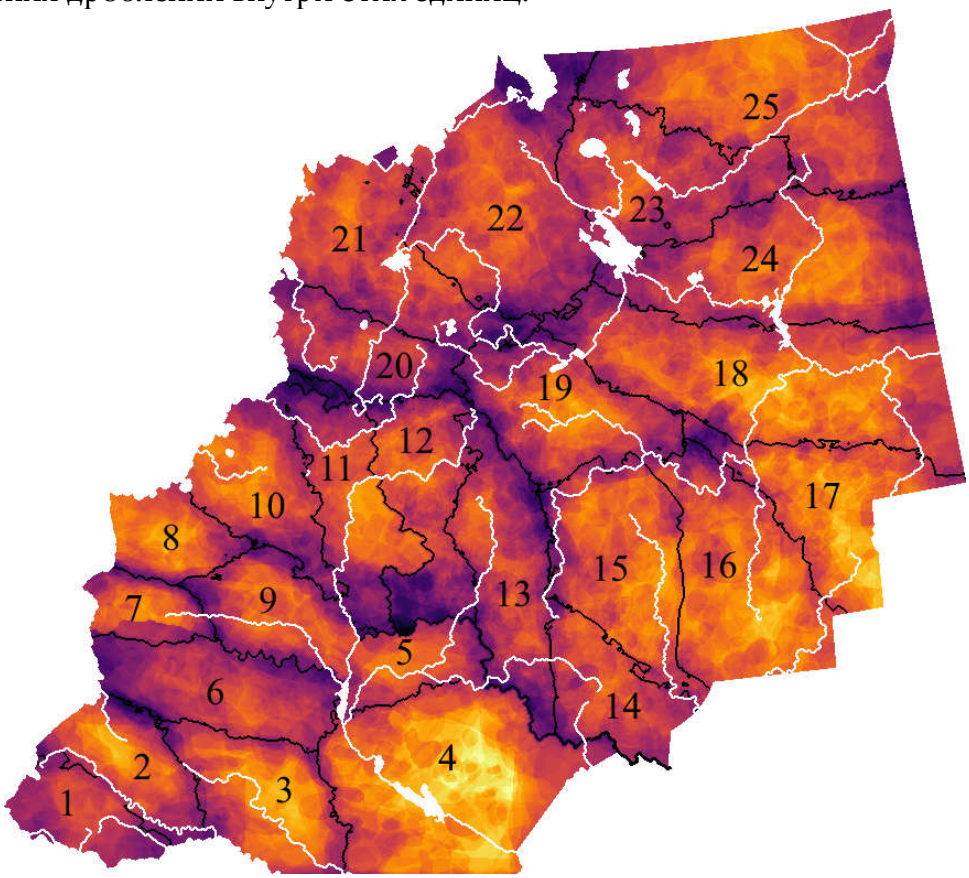

Рисунок 8. Дискретно-континуальный подход к классификации ВСЯ на 25 диалектных единиц.

При применении данного метода к классификации на 25 диалектных единиц самыми обширными темными участками на карте являются территория к западу от 19-го кластера (которая выделится в 26-й кластер), территория в центре Беларуси (выделится в 27-й кластер), территория на западе Украины на стыке 2-го и 6-го кластеров (28-й кластер), территории на востоке 24-го и 25-го кластеров (29-й и 30-й кластеры соответственно). Самый темный и при этом относительно небольшой участок на карте — юг 12-го кластера выделяется в отдельный 31-й кластер. Другие темные участки, в частности, территория к югу от 20-го кластера, территория на границе 13-го кластера, несмотря на свой цвет и размер выделяются в отдельные кластеры достаточно поздно, так как помимо оценки близости точек кластера к центру того кластера, к которому она отнесена, необходимо также учитывать, насколько далеко она находится от центров других кластеров.

Заключение

Данная работа — первый шаг в описании восточнославянского диалектного континуума с использованием компьютерных методов. В работе были взяты карты из сборника «Восточнославянских изоглосс» (ВСИ-1 1995; ВСИ-2 1998; ВСИ-3 2000; ВСИ-4 2006), были учтены диалектные явления, актуальные при классификации ВСЯ.

При сравнении полученных в этой работе диалектных членений на 5 и 25 диалектных единиц с предыдущими членениями (Захарова & Орлова 1970; Пшеничнова 1996; Аванесаў и др. 1969; Довгопол и др. 1977) наблюдается ряд сходств, границы наиболее крупных ареалов (например, севернорусское и южнорусское наречия) совпадают. При более дробном делении заметны различия, чаще всего объяснимые тем, что некоторые 2 единицы на нашей классификации уже распались, в то время как на сравниваемой карте все еще представляют единую общность и наоборот.

Также наше исследование позволяет не только назвать признаки, внесшие наибольший вклад в выделение кластера, но и измерить их вклад количественно. Так, при более дробных классификациях важную роль в выделении кластеров начинают играть лексические признаки, в то время как в классификации на 5 единиц признаки из всех разделов науки о языке оказывают влияние на формирование кластера.

Помимо дискретных диалектных членений в работе также предлагается континуальный подход к описанию восточнославянских языков. Он в некоторой мере обладает предсказательной силой, так как показывает переходные говоры, которые при более дробной классификации могут выделиться в новые диалектные единицы более низкого уровня.

В то же время уже сейчас заметны ограничения в исходном материале, из-за которых могут быть сделаны ложные выводы о диалектной принадлежности тех или иных, в особенности приграничных, населённых пунктов. В первую очередь это особенность собирания и внесения сведений о приграничных н./п. в ДАРЯ, АУМ, и ДАБМ. Однако благодаря большому объему данных неточность и неединообразность сбора, картографирования и последующего объединения сведений в “Восточнославянские изоглоссы” несущественно влияют на результат применения алгоритма. Для уточнения классификации необходимо пополнение базы данных новыми картами, а также внесение изменений в уже существующие карты, если они необходимы.

Литература

Аванесов, Р. И. (ред.). 1947. *Программа собирания сведений для составления диалектологического атласа русского языка*. Москва/Ленинград: Изд-во Академии наук СССР.

Аванесаў, Р. І., К. К. Атраховіч, Ю. Ф. Мацкевіч (ред.). 1969. *Лінгвістычная геаграфія і групоўка беларускіх гавораk (Карты)*. Минск: Акадэмія навук Беларускай ССР, Інстытут мовазнаўства імя Я. Коласа.

АУМ = Матвіяс, И. Г. и др. 1984–2001. *Атлас української мови*. Часть I-III. Київ.

ВСИ-1 = Попова, Т. В. (ред.). 1995. *Восточнославянские изоглоссы 1995*. Москва: Наука.

ВСИ-2 = Попова, Т. В. (ред.). 1998. *Восточнославянские изоглоссы. Вып. 2*. Москва: Наука.

ВСИ-3 = Попова, Т. В. (ред.). 2000. *Восточнославянские изоглоссы. Вып. 3*. Москва: Наука.

ВСИ-4 = Попова, Т. В. (ред.). 2006. *Восточнославянские изоглоссы. Вып. 4*. Москва: Наука.

ДАБМ = Аванесов, Р. И., Ю. Ф. Мацкевіч. 1963. *Дыялекталагічны атлас беларускай мовы*. Мінск: Выд-ва Акадэміі Навук БССР.

ДАРЯ = Аванесов, Р. И., С. В. Бромлей. 1986–2004. *Диалектологический атлас русского языка: Центр Европейской части СССР. Вып. I: Фонетика. Вып. II: Морфология. Вып. III: Часть 1. Лексика; Часть 2. Синтаксис и лексика*. Москва: Наука.

Довгопол, С. Ф., А. М. Залеський, Н. Н. Прилипко. 1977. *Говори української мови (збірник текстів)*. Київ: Академія наук Української РСР, Ордена Трудового Червоного Прапора інститут мовознавства ім. О. О. Потебні.

Дурново, Н. Н., Н. Н. Соколов, Д. Н. Ушаков (ред.). 1915. *Опыт диалектологической карты русского языка в Европе: с прил. очерка русской диалектологии*. *Труды Моск. Диалектол. Комис. вып. 5*. Москва: Синод. Тип.

Захарова, К. Ф., В. Г. Орлова. 1970. *Диалектное членение русского языка. Учеб. пособие для факультетов русского языка и литературы пед. ин-тов*. Москва: Просвещение.

Манусов, А. В., А. С. Кузьмина, Р. В. Ронько. 2024. *Банк диалектологических карт*. Доступно: https://dm.ruslang.ru/. Дата обращения 10.10.2024.

Пшеничнова, Н. Н. 1996. *Типология русских говоров*. Москва: Наука.

References

AUM = Matvijas, I. G. et al. 1984–2001. *Atlas ukrains'koj movy*. Parts I–III. Kyiv.

Avanesov, R. I. (ed.). 1947. *Programma sobiranija svedenij dl'a sostavlenija dialektologicheskogo atlasa russkogo jazyka*. Moskva / Leningrad: Izdatel'stvo Akademii nauk SSSR.

Avaniesaŭ, R. I., K. K. Atrachovič, Ju. F. Mackievič, Ju. F. (eds.). 1969. *Linguistic geography and groups of Belarusian dialects (Maps)*. Minsk: Institute of Linguistics of the National Academy of Sciences of Belarus.

DABM = Avanesov, R. I., Yu. F. Matskevich (eds.). 1963. *Dyyalektalagichny atlas belaruskaj movy*. Minsk: Vyd-va Akademii Navuk BSSR.

DARYA = Avanesov, R. I., S. V. Bromlej (eds.). 1986–2004. *Dialektologicheskij atlas russkogo yazyka: Tsentr Jevropejskoj chasti SSSR*. Moskva: Nauka.

Dovhopol, S. F., A. M. Zaleskyi, N. N. Prylypko. 1977. *Dialects of the Ukrainian language (collection of texts)*. Kyiv: Academy of Sciences of the Ukrainian Russian Socialist Republic, Potebnia Institute of Linguistics.

Durnovo, N. N., N. N. Sokolov, D. N. Ushakov (eds.). 1915. *Opyt dialektologicheskoj karty russkogo jazyka v Evrope: s pril. ocherka russkoj dialektologii*. *Trudy Mosk. Dialektol. Komis., vypusk 5.* Moskva: Sinod. Tip.

Hyvärinen, Aapo, Erkki Oja. 1997. A fast fixed-point algorithm for independent component analysis. *Neural Computation* 9(7): 1483–1492.

Hyvärinen, Aapo, Juha Karhunen, Erkki Oja. 2001. *Independent Component Analysis*. Hoboken: John Wiley & Sons.

Likas, Aristidis, Nikos Vlassis, Jakob Verbeek. 2003. The global k-means clustering algorithm. *Pattern Recognition* 36(2): 451–461.

Manusov, A. V., A. S. Kuzmina, R. V. Ronko. 2024. Collection of dialectological maps. Available at: http://emanresuaretne.pythonanywhere.com. Accessed 10.10.2024.

Marchenko, I. A. 2023. *Building A Classification of Russian Dialects Using Multidimensional Scaling*. Moscow: HSE University.

Pshenichnova, N. N. 1996. *Typology of Russian dialects*. Moscow: Nauka.

VSI-1 = Popova, T. V. (ed.). 1995. *Vostochnoslavyanskie izoglossy* 1995. Moscow: Nauka.

VSI-2 = Popova, T. V. (ed.). 1998. *Vostochnoslavyanskie izoglossy*. Vol. 2. Moscow: Nauka.

VSI-3 = Popova, T. V. (ed.). 2000. *Vostochnoslavyanskie izoglossy*. Vol. 3. Moscow: Nauka.

VSI-4 = Popova, T. V. (ed.). 2006. *Vostochnoslavyanskie izoglossy*. Vol. 4. Moscow: Nauka.

Zaharova, K. F., V. G. Orlova. 1970. *Dialectal division of the Russian language. A textbook for the faculty of Russian language and literature of pedagogical institutes*. Moscow: Prosvesh'enije.

Anastasiya Kuzmina, Arseniy Manusov. Dialectometric approach to the dialect classification of East Slavic languages based on the material of the collection "Vostochnoslavyanskie izoglossy"

The article proposes a new dialectometric approach to the division of East Slavic languages. Our dialectometry is based on the material from the collection of articles "Vostochnoslavyanskie izoglossy" ("East Slavic isoglosses", 1995–2006), which is a generalization of data from atlases of East Slavic languages (Dialectological atlas of the Russian language, Dialectological atlas of the Belarusian language, Atlas of the Ukrainian language). We trans-

ferred the VSI maps to electronic format, with each feature presented on the original map stored as a separate image. This procedure made it possible to construct a matrix of features contained in East Slavic atlases. The k-means method was used to construct clustering. New variants of dialect divisions of the East Slavic languages have been obtained, and significant dialectal features that are differential have been identified for the areas. This dialect division is compared with existing ones. The materials of "Vostochnoslavyanskie izoglossy" contain homogeneous data for all three languages, which allows to analyze the territory of the atlases as a single linguistic area and to study the East Slavic languages not only discretely, but also continuously. The article offers an analysis of cross-border dialects of the presented dialect division.

Keywords: dialectometry; quantitative methods in linguistics; East Slavic languages; Slavic languages; isoglosses; automatic classification; dialectology; geographical linguistics.