НАУЧНЫЙ ЖУРНАЛ
НАУКА И МИРОВОЗЗРЕНИЕ

УДК-81'28

# ИЗУЧЕНИЕ ДИАЛЕКТОВ С ПОМОЩЬЮ КОМПЬЮТЕРНЫХ МЕТОДОВ

**Новикова Валерия Дмитриевна**
Аспирант кафедры прикладной и математической лингвистики, Московский государственный университет имени М. В. Ломоносова
г. Москва, Россия

**Аннотация**

В настоящей масштабной фундаментальной исследовательской работе осуществляется развернутый, детальный, глубокий и многоаспектный научно-теоретический анализ методов компьютерной диалектологии, вычислительной лингвистики, геопространственного анализа и машинного обучения, применяемых для сбора, формализации, картографирования и исследования языковой вариативности в диалектах и говорах. В центре внимания находится исследование представления фонетических, лексических, морфологических и синтаксических признаков в виде структурированных матриц, графов и векторных пространств, методов кластеризации, многомерного шкалирования и диалектометрии, а также использования автоматического распознавания речи, корпусной обработки и нейросетевых моделей для анализа спонтанной диалектной речи. В статье подробно рассматриваются вопросы цифровой транскрибации и нормализации нестандартных орфографических и звуковых форм, пространственной интерполяции ареалов, влияния языковых контактов и миграций, а также этические, методологические и корпусные ограничения компьютерного анализа диалектов. На основе междисциплинарного синтеза диалектологии, вычислительной лингвистики, геоинформатики, статистики и машинного обучения формулируются концептуальные подходы к документированию, моделированию и сохранению диалектного разнообразия.

**Ключевые слова:** компьютерная диалектология, диалектометрия, вычислительная лингвистика, машинное обучение, автоматическое распознавание речи, пространственный анализ, корпусная лингвистика, языковая вариативность, кластеризация, диалектный атлас.

## Введение: трансформации диалектологии в цифровую эпоху

Диалектология на протяжении длительного времени опиралась на полевые экспедиции, ручное анкетирование носителей говоров, фонографическую или магнитофонную фиксацию речи и последующее составление печатных диалектологических атласов, словарей и описаний отдельных говоров. Подобные методы позволили накопить богатейший эмпирический материал о

пространственном распределении языковых явлений, исторических изоглоссах, фонетических переходах и лексическом разнообразии, однако при росте объемов данных, усложнении географических и социальных параметров и необходимости сопоставления множества признаков одновременно традиционные ручные процедуры сталкиваются с ограничениями по трудоемкости, субъективности интерпретации и масштабируемости.

Развитие вычислительной техники, цифровых средств аудио- и видеозаписи, геоинформационных систем, методов статистического анализа и машинного обучения открыло новые возможности для изучения диалектов. Компьютерные методы позволяют преобразовывать разнородные полевые наблюдения — транскрипции, аудиозаписи, словарные карточки, географические координаты населенных пунктов и социальные метаданные информантов — в единые машиночитаемые структуры данных, выявлять скрытые закономерности распределения признаков и моделировать пространственную и социальную динамику языковой вариативности.

Вместе с тем применение компьютерных методов в диалектологии требует строгого учета специфики живой разговорной и диалектной речи. Диалектные данные часто характеризуются отсутствием единой кодифицированной орфографии, высокой фонетической и морфологической вариативностью, неполнотой полевых выборок для отдельных населенных пунктов, фоновым шумом в архивных и полевых записях, а также постепенным размыванием традиционных говоров под влиянием литературного языка, средств массовой информации и урбанизации. Поэтому эффективное исследование диалектов с помощью вычислительных инструментов предполагает тесное взаимодействие лингвистов-диалектологов и специалистов по обработке данных.

Целью данного исследования является систематический научно-теоретический анализ компьютерных методов изучения диалектов, рассмотрение подходов к формализации и обработке диалектных данных, методов количественного и геопространственного анализа ареалов, технологий автоматической обработки звучащей речи и текстов, а также оценка методологических ограничений и перспектив развития компьютерной диалектологии.

**Формализованное представление данных и методы диалектометрии**

Основой компьютерного анализа диалектов является структурирование лингвистических признаков в форматах, пригодных для математической и алгоритмической обработки. В зависимости от задач исследования признаки могут относиться к фонетике и фонологии, лексике, словообразованию, морфологии или синтаксису, а единицами наблюдения выступают отдельные населенные пункты, регионы, конкретные информанты, диалектные тексты или аудиофрагменты. Для количественного сопоставления пунктов обследования строятся матрицы признаков, в которых строки соответствуют обследуемым точкам или информантам, столбцы — лингвистическим переменным или

вариантам, а ячейки — частотам, вероятностям употребления или закодированным категориальным значениям.

Важным направлением количественного анализа является диалектометрия — совокупность математических и статистических методов измерения лингвистических расстояний между диалектами и выявления структурных группировок на основе большого числа признаков. На базе матриц признаков вычисляются индексы сходства или различия между парами населенных пунктов, строятся матрицы расстояний и применяются методы многомерного анализа, включая иерархическую и неиерархическую кластеризацию, многомерное шкалирование и анализ главных компонент. Эти подходы позволяют оценивать степень близости говоров не по отдельным изолированным изоглоссам, а по совокупности признаков, выявляя переходные зоны, континуумы и кластеры относительной однородности.

Поскольку границы между диалектами редко бывают абсолютно резкими и часто носят градуальный характер, в компьютерной диалектологии применяются методы мягкой кластеризации, вероятностные модели и графовые представления, где узлы соответствуют населенным пунктам или вариантам, а ребра отражают степень сходства или интенсивность контактов. Сопоставление лингвистических расстояний с географическими дистанциями, транспортной связностью или историко-демографическими показателями дает возможность проверять гипотезы о пространственной диффузии инноваций, роли естественных природных барьеров и влиянии миграционных потоков на формирование диалектных ландшафтов.

**Геопространственное моделирование и цифровое картографирование ареалов**

Традиционная диалектология тесно связана с картографированием языковых явлений, а использование геоинформационных систем и методов пространственной статистики переводит диалектное картографирование на качественно новый уровень. Цифровые карты позволяют совмещать лингвистические данные с топографическими, гидрографическими, административными, демографическими и историческими слоями, анализируя пространственное распределение вариантов в непрерывном географическом пространстве.

На практике полевые обследования нередко охватывают лишь часть населенных пунктов региона, поэтому для оценки вероятности распространения признака в неизученных точках применяются методы пространственной интерполяции, включая кригинг, обратно взвешенные расстояния и пространственные авторегрессионные модели. Пространственная автокорреляция позволяет количественно оценить, насколько значения лингвистических признаков в близлежащих пунктах зависят от географической близости, и выделить области статистически значимой концентрации или контраста вариантов.

Компьютерное картографирование также облегчает визуализацию агрегированных диалектометрических показателей, таких как плотность пучков изоглосс, локальная вариативность или степень отклонения говора от литературной нормы или регионального эталона. При этом важно учитывать искажения, возникающие при механическом проведении границ по дискретным точкам выборки, и сопровождать карты оценками неопределенности и плотности наблюдений, чтобы визуализация отражала реальную структуру данных, а не артефакты интерполяции.

**Автоматическая обработка речи, транскрибация и корпусный анализ**

Значительная часть диалектологического материала представлена устной речью, зафиксированной на аудио- и видеоносителях, а также текстами в нестандартной орфографии, диалектными словарями и записями в социальных сетях или локальных сообществах. Обработка такого материала с помощью компьютерных методов требует решения задач автоматического распознавания речи, сегментации, фонетического анализа и нормализации текстов. Стандартные системы распознавания речи, обученные преимущественно на нормативной литературной речи и больших объемах стандартных аудиоданных, часто испытывают трудности при распознавании диалектной фонетики, специфической интонации, редукции гласных, особых морфологических форм и региональной лексики.

Для адаптации акустических и языковых моделей к диалектному материалу применяются методы дообучения на специализированных диалектных корпусах, кросс-лингвального и междиалектного переноса знаний, а также синтеза обучающих данных с моделированием фонетических вариаций. Автоматическая сегментация и выравнивание звукового сигнала с транскриптами позволяют ускорять подготовку корпусов устной речи, выделять отдельные словоформы и проводить акустический анализ параметров — таких как формантные частоты гласных, длительность звуков или контуры основного тона — для объективного сопоставления произносительных особенностей различных говоров.

При работе с письменными или транскрибированными диалектными текстами ключевую роль играют методы корпусной лингвистики и обработки естественного языка. Поскольку диалектные написания могут варьировать даже в пределах одного текста или одного информанта, применяются алгоритмы нормализации и лемматизации, частотно-статистический анализ, выявление коллокаций и морфосинтаксическая разметка, адаптированная к нестандартным формам. Это позволяет исследовать употребление диалектных лексем, грамматических конструкций и дискурсивных маркеров в реальном контексте, а также сопоставлять диалектные корпуса с корпусами литературного языка и региональных койне.

## Машинное обучение, нейросетевые модели и проблемы интерпретируемости

В последние годы в изучении диалектов активно используются методы машинного обучения и глубокого обучения, способные работать с высокоразмерными признаками, нелинейными зависимостями и неструктурированными данными, включая звуковые сигналы и контекстные последовательности слов. Задачи классификации диалектов или регионов по тексту или речи, предсказания географического происхождения автора, выявления диалектных маркеров и моделирования языковой вариативности решаются с помощью линейных и ансамблевых моделей, рекуррентных и сверточных нейронных сетей, а также трансформерных архитектур и самообучаемых речевых представлений.

Самообучаемые модели представления речи и языка позволяют извлекать информативные скрытые признаки из больших массивов неразмеченных данных, что особенно ценно для малоресурсных диалектов и говоров, где объем ручной разметки ограничен. Векторные представления слов, контекстных фрагментов или речевых сегментов дают возможность измерять семантическую, фонетическую или стилистическую близость вариантов и отслеживать тонкие градации между стандартными и диалектными формами.

Вместе с тем применение сложных нейросетевых моделей в диалектологии требует особого внимания к интерпретируемости и смещениям в данных. Если модель классифицирует говор по косвенным признакам — например, по упоминанию местных топонимов, профессиональных реалий, шумовому фону записи или демографическим перекосам в обучающей выборке — она может демонстрировать высокую точность на тестовых данных, не опираясь на подлинные системные языковые различия. Поэтому в лингвистических исследованиях важно сочетать прогностическую мощность моделей с методами анализа важности признаков, проверкой на кросс-региональных или кросс-информантных разбиениях и содержательной лингвистической верификацией полученных результатов.

## Методологические ограничения, этика данных и перспективы развития

Несмотря на расширяющиеся возможности вычислительных методов, компьютерная диалектология сталкивается с рядом устойчивых методологических и практических вызовов. Существенная часть исторических и полевых материалов неоднородна по качеству фиксации, полноте охвата территории и сопоставимости анкет разных эпох, а оцифровка и стандартизация архивных коллекций требуют значительных ресурсов. Кроме того, языковая ситуация в многих регионах динамично меняется: традиционные сельские диалекты взаимодействуют с городскими социолектами, региональными вариантами языка и цифровой коммуникацией, что усложняет четкое разграничение дискретных диалектов и требует учета социальных, возрастных и коммуникативных факторов.

Важное значение имеют этические и правовые аспекты работы с диалектными данными, включая получение информированного согласия информантов, защиту персональных данных при публикации аудиозаписей, геометок и биографических сведений, а также уважительное отношение к языковым сообществам и сохранение их культурного наследия. Открытые стандарты метаданных, воспроизводимость вычислительных экспериментов и документирование этапов предобработки данных становятся необходимым условием научной достоверности исследований.

## Заключение

Изучение диалектов с помощью компьютерных методов представляет собой динамично развивающееся междисциплинарное направление, объединяющее традиционную диалектологию, вычислительную лингвистику, статистику, геоинформатику и машинное обучение. Формализация лингвистических признаков, диалектометрический анализ, цифровое картографирование, автоматическая обработка речи и текстов, а также нейросетевое моделирование расширяют возможности исследования пространственной и социальной вариативности языка на больших массивах данных.

Эффективное применение вычислительных подходов требует критического учета особенностей диалектного материала, неполноты выборок, изменчивости языковой среды и необходимости содержательной лингвистической интерпретации результатов. Сочетание современных цифровых технологий с глубокой лингвистической экспертизой и ответственными принципами работы с данными создает прочную основу для документирования, анализа и сохранения диалектного наследия.

## Литература


1. Касаткин Л. Л. Русская диалектология. — М.: Академия, 2020. — 288 с.
2. Баранов А. Н. Введение в прикладную лингвистику. — М.: URSS, 2020. — 360 с.
3. Захаров В. П., Богданова С. Ю. Корпусная лингвистика. — СПб.: СПбГУ, 2020. — 234 с.
4. Копотев М. В., Сметанина М. П. Введение в компьютерную лингвистику. — СПб.: Златоуст, 2021. — 256 с.
5. Крылов С. А. Количественные методы в лингвистике: учебное пособие. — М.: Флинта, 2019. — 240 с.
6. Яковлева Е. В., Морозова Н. А. Геоинформационные системы и пространственный анализ в гуманитарных исследованиях. — Новосибирск: СГУГиТ, 2021. — 216 с.
7. Nerbonne J., Kretzschmar W. A. Dialectometry and Computational Dialectology. — Literary and Linguistic Computing. — 2019. — Vol. 28, No. 4. — P. 625–636.