



КОРПУСНЫЕ МЕТОДЫ ИССЛЕДОВАНИЯ МОРФОЛОГИЧЕСКИХ КАТЕГОРИЙ

Мавлянова Тамилла Бахрилловна

*д.ф.ф.н. (PhD) кафедры русского языка и литературы
Самаркандского государственного педагогического института;*

Зайниддиновна Нигина

*студентка 3 курса факультета языков
Самаркандского государственного педагогического института*

Аннотация: Настоящая статья посвящена разработке методологических принципов корпусного анализа морфологических категорий, рассматриваемых не как статичные классификаторы, а как динамические, вероятностные феномены, чье реальное функционирование раскрывается лишь через анализ больших массивов текстов. В работе обосновывается переход от структурно-семантического описания грамматических форм к количественному анализу их дистрибутивных свойств — частотности, коллокационной сочетаемости и контекстуальной обусловленности. На материале современного русского языка демонстрируется, как корпусные методы позволяют выявить зоны грамматической вариативности (например, конкуренция падежных форм или колебания в выборе видовременных форм глагола), а также проследить микродиахронические сдвиги в морфологической системе, фиксируя процессы грамматикализации и прагматикализации. Особое внимание уделяется проблеме интерпретации количественных данных: частотность признается значимым, но не единственным критерием нормативности и системной релевантности, что требует разработки комплексных методов, сочетающих корпусные данные со сведениями о типологической дистрибуции явлений и их психолингвистической реальности. Статья адресована лингвистам, интересующимся применением эмпирических методов в грамматических исследованиях и проблемами современной морфологии.

Ключевые Слова: корпусная лингвистика, морфология, грамматическая вариативность, количественные методы.

Переход к большим лингвистическим данным, ознаменовавший развитие науки о языке на рубеже XX–XXI веков, привел к фундаментальной трансформации исследовательских практик в области морфологии. Традиционная грамматика, опиравшаяся преимущественно на интроспекцию исследователя и анализ ограниченного корпуса



иллюстративных примеров из художественной литературы, столкнулась с необходимостью пересмотра собственных оснований, когда выяснилось, что интуитивные представления о грамматической правильности и нормативности зачастую не подтверждаются данными о реальном функционировании языковых единиц. Этот эпистемологический сдвиг, осуществленный в рамках корпусной лингвистики окончательно сформировавшейся как самостоятельное направление в последней четверти двадцатого столетия, позволяет сегодня рассматривать морфологическую систему языка как вероятностную структуру, где грамматические категории проявляют себя не столько через жесткие императивные правила, сколько через статистически значимые паттерны дистрибуции, коллокационные предпочтения и контекстуальную обусловленность выбора той или иной формы.¹³⁶ Именно это понимание морфологии как области постоянной вариативности и конкуренции форм, где канонические инварианты сосуществуют с многочисленными периферийными реализациями, составляет теоретическую базу для применения корпусных методов, которые, в отличие от традиционных подходов, оперирующих логическими категориями, позволяют наблюдать языковые явления в их естественной среде обитания — в реальных текстах, порожденных носителями языка.

Исторически сложилось так, что в фокусе внимания корпусной морфологии оказались категории, характеризующиеся высокой степенью формальной и семантической вариативности. Прежде всего это касается тех областей грамматической системы, где наблюдаются процессы конкуренции между синонимичными или функционально близкими формами, например между окончаниями родительного падежа единственного числа мужского рода, где варианты на -а/-я и -у/-ю демонстрируют сложное взаимодействие лексических и грамматических факторов, или между формами именительного и родительного падежа при отрицании, где выбор падежной формы прямого объекта определяется целым комплексом условий, включающих аспектуальную характеристику глагола, одушевленность существительного, а также коммуникативную структуру высказывания. Именно корпусные исследования позволили выйти за пределы утверждений о факультативной вариативности и предложить квантитативно обоснованные модели, объясняющие распределение форм в зависимости от частотности существительного, типа отрицания и синтаксической конструкции в целом. В этом

¹³⁶ Баранов А.Н., Плунгян В.А., Рахилина Е.В. Путеводитель по дискурсивным словам русского языка. М.: Помовский и партнеры, 2007. С. 12–15.



отношении особенно показательными являются работы, выполненные на материале Национального корпуса русского языка, где исследователям удалось продемонстрировать, что для высокочастотных абстрактных имен, таких как, например, лексема, обозначающая интеллектуальный процесс 'смысл', форма родительного падежа с окончанием -у практически вытеснена более нейтральным окончанием -а, тогда как в сфере вещественных существительных, называющих конкретные вещества и материалы, указанная флексия сохраняет свою продуктивность, особенно при обозначении меры или части целого.¹³⁷

Теоретическая значимость подобных наблюдений выходит далеко за пределы описания конкретных грамматических фактов, поскольку они позволяют по-новому осмыслить само устройство морфологического компонента языка. Традиционная морфология, находившаяся под сильным влиянием структурализма с его стремлением к бинарным оппозициям и дискретным категориям, долгое время рассматривала грамматическую форму как инвариантный знак, чье значение и функции могут быть описаны независимо от конкретного лексического наполнения и контекста употребления. Однако корпусные исследования демонстрируют, что морфологические категории не являются гомогенными и единообразными: разные лексемы, принадлежащие к одному и тому же грамматическому классу, могут демонстрировать различное поведение при взаимодействии с грамматическими правилами.

Данный феномен, получивший в литературе название "лексически специфической грамматики", заставляет нас пересмотреть представления о соотношении лексикона и грамматики, рассматривая их не как автономные модули, а как континуум, где лексические единицы обладают собственными грамматическими профилями, то есть предпочтениями в отношении морфологических форм и конструкций, в которых они предпочитают употребляться. В этой связи можно указать на исследования глагольной аспектуальности в русском языке, где корпусный анализ позволил выявить значительные различия в видовой парности для разных глагольных лексем. Оказалось, что частотность употребления совершенного и несовершенного вида для одного и того же глагола может варьироваться в зависимости от семантической зоны: глаголы, обозначающие однонаправленное физическое действие, демонстрируют

¹³⁷ Савчук С.О., Архангельский Т.А., Бонч-Осмоловская А.А., Донина О.В., Кузнецова Ю.Н., Ляшевская О.Н., Орехов Б.В., Подрядчикова М.В. Национальный корпус русского языка 2.0: новые возможности и перспективы развития // Вопросы языкознания. 2024. № 2. С. 7–34. (конкретно с. 18–22)



иную видовую дистрибуцию по сравнению с глаголами интеллектуальной деятельности или глаголами, описывающими изменения состояния .

Методологически важным представляется то обстоятельство, что современные корпусные исследования русской морфологии опираются не только на письменные тексты, но все в большей степени используют данные устных корпусов и корпусов спонтанной речи, что открывает принципиально новые перспективы для анализа грамматических категорий. В устной речи, в отличие от кодифицированных письменных текстов, процессы морфологической вариативности проявляются с максимальной очевидностью, поскольку здесь ослаблен контроль нормы и наблюдается значительно более высокая степень спонтанности. Изучение мультимедийного русского корпуса, которым располагает Национальный корпус русского языка, позволяет исследовать такие феномены, которые практически не фиксируются в письменных текстах, например, случаи оговорок, морфологических ошибок, самокоррекций и гибридных форм, возникающих в процессе речепорождения. Особый интерес в этом контексте представляют исследования, посвященные функционированию дискурсивных слов и прагматических маркеров, возникающих в результате грамматикализации полнозначных лексем и демонстрирующих утрату первоначальных морфологических характеристик. Показательным примером является история формирования дискурсивного маркера, который восходит к глагольной форме первого лица множественного числа будущего времени, исходно указывавшей на будущее речевое действие, но в синхронном употреблении утратившей эту семантику и превратившейся в оператор, вводящий пояснение или пример, где говорящий снимает с себя ответственность за точность номинации.¹³⁸ Аналогичные процессы прагматикализации, доступные для наблюдения именно благодаря корпусным данным, затрагивают целый ряд грамматических форм, включая императивные и инфинитивные конструкции, постепенно приобретающие модальные и оценочные значения.

Одним из наиболее перспективных направлений современной корпусно-ориентированной морфологии является микродиакронический анализ, позволяющий проследить изменения в грамматической системе языка на относительно коротких временных отрезках от пятидесяти до ста лет. В отличие от классической исторической лингвистики, которая оперирует столетиями и даже тысячелетиями, микродиакронический

¹³⁸ Подлесская В.И. Скажем так: дискурсивные маркеры, восходящие к глаголам речи в русском языке // Вопросы языкознания. 2024. № 2.



подход, ставший возможным благодаря созданию представительных временных корпусов, охватывающих период с XVIII по XXI век, позволяет фиксировать процессы грамматических изменений в режиме реального времени, наблюдая за тем, как новые формы проникают в язык, конкурируют со старыми и постепенно либо вытесняют их, либо занимают свою нишу в функционально-стилистическом пространстве языка. В русистике применение этого метода привело к получению целого ряда нетривиальных результатов, существенно корректирующих представления о темпах и механизмах языковой эволюции. В частности, исследования показали, что многие морфологические изменения, считавшиеся завершёнными ещё в XVIII–XIX веках, на самом деле продолжают и в современную эпоху, хотя и в менее заметных формах. Классический пример здесь — история притяжательного местоимения, образованного от основы третьего лица множественного числа, с суффиксом, не свойственным литературной норме, которое на протяжении двухсот лет продолжает попытки проникнуть в литературный язык из просторечия и диалектов. Данные диахронического корпуса позволяют проследить, как частотность этой формы то возрастала, то снижалась в зависимости от экстралингвистических факторов, включая процессы демократизации языка и ослабления кодификаторской деятельности, и какую роль в её маргинализации сыграла школьная грамматическая традиция.

В этом же ряду стоит отметить исследования, посвященные трансформации русской деепричастной системы. Исходно, в древнерусский период, деепричастия образовывались от глаголов любого вида с помощью суффиксов, различавшихся по временному значению, однако к XIX–XX векам система существенно перестроилась: суффикс, исходно маркировавший прошедшее время, стал закрепляться за деепричастиями совершенного вида, в то время как суффикс, связанный с настоящим временем, стал доминировать в сфере несовершенного вида.

Этот процесс, однако, не является завершённым до сих пор: корпусные данные фиксируют значительное количество деепричастий, образованных по "неправильным" моделям, особенно в разговорной речи и публицистических текстах, что свидетельствует о продолжающейся перестройке этого фрагмента грамматической системы. Более того, микродиахронический анализ позволяет увидеть, что грамматическая норма фиксирует далеко не все возможные варианты и что реальная языковая практика значительно разнообразнее предписаний грамматик и словарей. Эта множественность грамматических вариантов, многие из которых находятся в разных стадиях нормативного одобрения или



отвержения, составляет основной объект исследования корпусной морфологии.

Следует особо подчеркнуть, что становление корпусной морфологии как самостоятельной исследовательской области было бы невозможно без параллельного развития методов автоматической обработки текстов и создания систем многоуровневой лингвистической аннотации. Морфологическая разметка, или тэггинг, — это процесс приписывания каждому слову в тексте его грамматических характеристик, таких как часть речи, род, число, падеж, время, лицо и так далее. Без такой разметки сколько-нибудь систематический анализ грамматических явлений на корпусном материале был бы практически невозможен, поскольку исследователь физически не способен вручную обработать миллионы словоупотреблений. Именно разработка надежных морфологических парсеров, способных автоматически снимать грамматическую омонимию и приписывать слову корректный грамматический ярлык, создала технологическую базу для корпусных исследований. В случае с русским языком эта задача оказалась особенно сложной в силу его высокоразвитой флективной морфологии, где одна и та же словоформа может реализовывать несколько различных грамматических значений, а также множественности парадигм склонения и спряжения. Создание системы морфологической разметки для Национального корпуса русского языка, осуществленное в начале двухтысячных годов, стало важной вехой не только для русистики, но и для мировой компьютерной лингвистики, продемонстрировав возможности автоматического анализа грамматики высокосинтетического языка.

Важно, однако, понимать, что автоматическая морфологическая разметка не является полностью точной и содержит определенный процент ошибок, который варьируется в зависимости от жанра текста, степени омонимии в конкретных словоформах и качества самого парсера. Поэтому методологически грамотное корпусное исследование предполагает не только использование автоматической разметки, но и верификацию полученных данных ручными методами, особенно в тех случаях, когда анализируются редкие или нетипичные морфологические явления. Более того, автоматическая разметка фиксирует лишь то, что заложено в ее алгоритмах, и не способна выявлять принципиально новые или не описанные в грамматиках явления, если они не соответствуют существующим категориальным схемам. Это означает, что корпусные методы должны сочетаться с традиционными качественными методами лингвистического анализа, поскольку сам по себе подсчет частотностей и дистрибуций, сколь бы точен он ни был, еще не дает понимания того,



почему грамматическая система устроена именно таким образом и какие механизмы лежат в основе наблюдаемых распределений.

Интерпретация количественных данных — это, пожалуй, самая сложная теоретическая проблема современной корпусной морфологии. Выявление статистически значимых различий в частоте употребления тех или иных грамматических форм не является самоцелью, но лишь исходным пунктом для лингвистического анализа, который должен объяснить эти различия, связав их с семантическими, прагматическими или когнитивными факторами. Существует методологическая опасность подмены лингвистического объяснения простой констатацией количественных фактов: утверждение о том, что одна форма встречается в корпусе чаще, чем другая, само по себе не является лингвистически содержательным, если мы не понимаем, почему это так.

Именно поэтому в последнее время все большее распространение получают методы многомерного статистического анализа, такие как множественная регрессия, анализ соответствий и логистическая регрессия, которые позволяют одновременно учитывать влияние нескольких факторов на выбор грамматической формы. Эти методы хорошо зарекомендовали себя при анализе падежной вариантности при отрицании, где выбор между родительным и винительным падежом прямого объекта определяется не одним фактором, а их сложным взаимодействием, и только статистический анализ позволяет оценить относительный вклад каждого из этих факторов и предсказать вероятность выбора той или иной формы в конкретном контексте.

Кроме того, последние годы ознаменовались стремительным развитием методов корпусной семантики, которые все активнее проникают в морфологические исследования. Речь идет о таких понятиях, как грамматический профиль лексемы, то есть распределение грамматических форм для данной лексической единицы по сравнению с другими единицами того же класса, и коллексема, то есть статистически значимая связь между грамматической формой и конкретными лексическими единицами, которые предпочитают эту форму. Анализ коллексем позволяет выявить семантические классы лексем, характеризующиеся сходным грамматическим поведением, и тем самым проверить гипотезы о семантической мотивированности грамматических правил. Применительно к русской морфологии, например, было показано, что для родительного падежа при отрицании существует четкая коллексемная структура: одни глаголы с высокой вероятностью требуют родительного падежа, другие — винительного, а третьи демонстрируют вариативность, причем эта дистрибуция коррелирует со степенью



воздействия глагола на объект, его определенностью и референтной значимостью. Подобные результаты, полученные на строго количественной основе, существенно углубляют наши представления о природе грамматической категории падежа и о том, как конкретные лексические единицы взаимодействуют с абстрактными грамматическими правилами.

Нельзя не отметить, что корпусные методы открывают новые возможности для изучения морфологических категорий в типологической перспективе. Создание параллельных корпусов, содержащих оригинальные тексты на одном языке и их переводы на другие языки, позволяет сопоставлять способы выражения грамматических значений в разных языковых системах, выявлять универсальные и идиоэтнические черты в устройстве морфологии. Для русского языка особенно актуальными являются контрастивные исследования, сопоставляющие его грамматические категории с категориями германских, романских и тюркских языков, поскольку такие сопоставления помогают лучше понять специфику именно русской грамматической системы, например, что именно отличает русскую систему глагольного вида от аналогичных систем в других славянских языках или в языках неславянских. В этом отношении интерес представляет типологическое исследование паремиологического фонда, включающее сопоставительный анализ морфологических характеристик соматических лексем в русских и других национальных пословицах. Как показывают такие исследования, морфологические модели, представленные в поговорках, не всегда соответствуют общезыковым тенденциям, что свидетельствует о том, что паремиологический фонд консервирует более архаические грамматические формы, которые в свободном употреблении уже ушли из языка или стали периферийными.

Помимо синхронных и диахронических исследований, корпусные методы все активнее применяются для анализа морфологии в ее связи с когнитивными процессами освоения языка, особенно в контексте изучения русского языка как неродного и как эритажного. Русский учебный корпус, включающий тексты, порожденные изучающими русский язык, предоставляет уникальную возможность наблюдать за тем, как происходит усвоение морфологических категорий, какие трудности возникают у носителей разных родных языков, и в каком порядке осваиваются грамматические формы.

В частности, исследования, основанные на этом ресурсе, показывают, что система падежей и система глагольного вида представляют собой особенно сложные зоны для усвоения, причем ошибки, совершаемые



изучающими, отнюдь не хаотичны, а подчиняются определенным закономерностям, отражающим универсальные стратегии упрощения грамматической системы. Особый интерес представляют исследования эритажных носителей, то есть тех, кто освоил русский язык в детстве в условиях билингвизма, но доминирующим для них стал другой язык. В речи таких носителей наблюдаются феномены языковой аттриции, то есть утраты или изменения грамматических компетенций, причем эти изменения носят не случайный, а системный характер: прежде всего, как показывают корпусные данные, утрачиваются прежде всего те морфологические категории и формы, которые обладают низкой частотностью, низкой функциональной нагрузкой или не имеют однозначных соответствий в доминирующем языке носителя.

Таким образом, корпусная морфология представляет собой динамично развивающуюся область лингвистики, методологический аппарат которой позволяет по-новому взглянуть на проблемы, традиционно считавшиеся центральными для грамматической теории. Признание вариативности и вероятностного характера грамматических явлений, выявление семантических и прагматических факторов, управляющих выбором грамматической формы, анализ микро- и нанодиахронических сдвигов в морфологической системе — все это становится возможным благодаря обращению к большим массивам аутентичных текстов, обработанных с помощью систем многоуровневой лингвистической аннотации. Вместе с тем, развитие корпусных методов ставит перед исследователями и новые теоретические вызовы, главный из которых связан с необходимостью интеграции количественного и качественного подходов. Простое накопление статистических данных, сколь бы впечатляющими по объему они ни были, не должно становиться самоцелью; истинная задача лингвиста — объяснить, почему распределение грамматических форм оказывается именно таким, а не иным, и какие именно свойства языковой системы и речевой деятельности продуцируют наблюдаемые паттерны. Иными словами, корпусные данные должны быть не конечной точкой исследования, но отправной точкой для лингвистической теории, которая способна не только описывать, но и прогнозировать грамматическое поведение языковых единиц, а также объяснять механизмы, лежащие в основе грамматических изменений.

Представляется, что будущее корпусной морфологии лежит в области интеграции нескольких методологических подходов. Во-первых, необходимо дальнейшее совершенствование систем автоматической аннотации, которые должны стать более точными и учитывать не только морфологические, но и синтаксические, семантические и даже



прагматические характеристики высказывания. Во-вторых, актуальным является создание специализированных корпусов, ориентированных на решение конкретных морфологических проблем, например, корпусов сбалансированных по жанрово-стилистическим параметрам для изучения функционально-стилистической дифференциации грамматических форм, или корпусов раннего детского усвоения для верификации гипотез о порядке освоения грамматических категорий. В-третьих, необходимо дальнейшее развитие методов статистического анализа, которые позволили бы не только описывать, но и визуализировать сложные многомерные данные, делая результаты корпусных исследований наглядными и интерпретируемыми для широкого круга лингвистов. В-четвертых, и это, пожалуй, самое важное, следует продолжать работу по теоретическому осмыслению статуса корпусных данных в системе лингвистического знания: каков онтологический статус частотностных данных, в какой мере частотность отражает грамматичность, и возможна ли корпусная верификация универсальных грамматических принципов, не сводящихся к дистрибутивным характеристикам. Только при условии позитивного решения этих теоретических вопросов корпусные методы смогут занять то место в иерархии лингвистических методов, которое соответствует их реальному эвристическому потенциалу — не как замена традиционным методам анализа, но как их важнейшее дополнение и расширение.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Нурутдинова А.Р. Синтаксическое и морфологическое «поведение» соматических лексем в пословицах: корпусное исследование в разноструктурных языках // Русская грамматика: полипарадигмальность как методологический принцип современных научных исследований. Иркутск: Иркутский государственный университет, 2025. С. 515–519.
2. Ляшевская О.Н. Корпусные инструменты в грамматических исследованиях русского языка. М.: Издательский дом ЯСК, Рукописные памятники Древней Руси, 2016. 520 с.
3. Савчук С.О., Архангельский Т.А., Бонч-Осмоловская А.А., Донина О.В., Кузнецова Ю.Н., Ляшевская О.Н., Орехов Б.В., Подрядчикова М.В. Национальный корпус русского языка 2.0: новые возможности и перспективы развития // Вопросы языкознания. 2024. № 2. С. 7–34.
4. Подлесская В.И. Скажем так: дискурсивные маркеры, восходящие к глаголам речи в русском языке // Вопросы языкознания. 2024. № 2.



5. Gries, S. Th., Wulff, S., & Davies, M. (Eds.). *Corpus-Linguistic Applications: Current Studies, New Directions*. Amsterdam & New York: Rodopi, 2010.
7. Добрушина Е.Р. Корпусные методы как инструмент исследования микродиахронии русского языка XVIII–XXI веков: Автореферат диссертации на соискание ученой степени доктора филологических наук. М.: НИУ ВШЭ, 2022.