ВЕСТНИК РОССИЙСКОЙ АКАДЕМИИ НАУК, 2024, том 94, № 9, с. 787–794

С КАФЕДРЫ ПРЕЗИДИУМА РАН

КОРПУСНАЯ ЛИНГВИСТИКА НА СОВРЕМЕННОМ ЭТАПЕ

© 2024 г. В.А. Плунгян[a,b,*]

[a]*Институт русского языка им. В.В. Виноградова РАН, Москва, Россия*
[b]*Институт языкознания РАН, Москва, Россия*

**E-mail: plungian@iling-ran.ru*

Поступила в редакцию 06.08.2024 г.
После доработки 06.08.2024 г.
Принята к публикации 12.08.2024 г.

Даётся общее представление о корпусной лингвистике, её истории, методах и влиянии на современные представления об изучении языка, которое обычно обозначается как “корпусная революция”.

Ключевые слова: корпусная лингвистика, теория языка, лингвистическая методология, функциональная лингвистика.

DOI: 10.31857/S0869587324090018, **EDN:** FCHMFE

Цель настоящих заметок – дать краткую характеристику такого сравнительно нового направления науки о языке, как корпусная лингвистика, коснувшись обстоятельств её возникновения и постепенной интеграции в центральные области теории языка, а также того влияния, которое она оказывает на современные представления об устройстве языка и задачах лингвистики в целом. В своё время нам уже доводилось писать о “корпусной революции” в лингвистике [1, 2]; прошедшие годы подтвердили эти тенденции – и то, что полтора десятка лет назад кому-то, может быть, казалось спорным и неочевидным, сегодня уже не вызывает возражений у подавляющего большинства исследователей. По поводу важнейшей роли корпусов и корпусных методов в изучении языка у лингвистов наших дней имеется, как принято говорить, консенсус.

Напомним в начале, что, собственно, понимается под корпусом языка. Лингвисты обычно говорят, что корпус определённого языка – это сложно устроенная “информационно-справочная система”, основанная на оцифрованном собрании текстов этого языка. Таким образом, корпус изначально противопоставляется простой коллекции (или “библиотеке”) текстов – и в том числе электронной коллекции – тем, что корпус является специальным образом обработанной коллекцией текстов, снабжённой особым научным аппаратом. Он обычно называется корпусной разметкой, или аннотацией, и именно аннотация и составляет важнейшую часть корпуса, определяя его специфику, его отличительные черты и его эффективность в качестве исследовательского инструмента. Грубо говоря, аннотация – это та информация об элементах текста (и тексте в целом), которая может быть интересна исследователю и которая может быть автоматически извлечена из корпуса при применении специальных компьютерных программ поиска[1]. Существенно, что

ПЛУНГЯН Владимир Александрович – академик РАН, заместитель директора по научной работе ИРЯ РАН, заведующий сектором типологии ИЯз РАН.

[1] Такие программы обычно называются “поисковой системой” или, на программистском жаргоне, “поисковым движком” (англ. search engine); поиск осуществляется по текстам, превращённым в базу данных, и учитывает все параметры аннотации (которая предварительно вносится в корпус или автоматически, или, в сложных случаях, вручную). Достижение оптимальной быстроты и эффективности существующих поисковых систем – особая проблема современных компьютерных технологий автоматической обработки текста (связанных далеко не только с корпусными задачами). Об этой стороне корпусной лингвистики мы далее специально говорить не будем; беглый обзор проблем, доступный в том числе и не специалисту, можно найти, например, в статье [3].

аннотация вносится в корпус для того, чтобы по её параметрам мог вестись поиск нужной информации (что и делает корпус "справочно-информационной системой"); поиск осуществляется не только по всем параметрам аннотации, но и по всем их возможным комбинациям. Если учитывать, что при аннотации отмечаются не только грамматические признаки элементов текста (слов, или, точнее, словоформ – впрочем, в корпусной лингвистике для отдельной самостоятельной единицы текста даже более принят термин "словоупотребление"), но и так называемые мета-признаки текста в целом (автор, дата создания, жанр, тематика и т.п.), то становится понятно, насколько гибкими и многообразными могут быть поисковые запросы к корпусу с хорошей аннотацией. Например, легко представить себе запрос такого вида: «Все художественные тексты И.С. Тургенева, написанные с 1850 по 1880 годы и содержащие сочетание слова *"роза"* и прилагательного цвета»[2].

Одним из важнейших артефактов корпусной лингвистики является так называемый "национальный корпус" языка. Это сочетание является вполне строгим термином – таким образом обозначается сбалансированное собрание текстов на некотором языке, представляющее этот язык с максимальной полнотой на всём пространстве и времени его существования. Иными словами, в национальный корпус входит максимально возможное число текстов, созданных на данном языке и в тех жанровых пропорциях, которые характерны для их бытования в языке в целом; тем самым, национальный корпус призван максимально адекватно отразить все те явления, которые в данном языке возможны, причём их количественная представленность в корпусе должна отражать их реальную частотность в языке. Подобного рода система должна обеспечивать все виды научных исследований языка (в том числе диахронические исследования, так как в языках с длительной письменной историей национальный корпус включает тексты всех исторических периодов).

Сам термин "национальный корпус" возник, как известно, ещё в 1990-е годы – первоначально как обозначение проекта Британского национального корпуса (British National Corpus, BNC); в названии подчёркивалась специфика британского варианта английского языка, который данным корпусом документировался. Поскольку BNC был первым электронным корпусом, претендовавшим на представительность и полноту и задавшим своего рода стандартный формат для таких систем, термин "национальный" стал использоваться и для обозначения других крупных представительных корпусов, которые стали появляться во второй половине 1990-х годов и позднее – американского английского, чешского, польского, греческого и др. [4]. Большинство из этих языков имело только один стандартный вариант, так что прилагательное "национальный" в их обозначении не имело того специфического смысла, которое оно выражало в составе корпуса BNC; фактически, оно превратилось в синоним прилагательного "представительный" (или даже просто "большой"), но в этом качестве – то есть для обозначения некоторого "главного" корпусного ресурса для определённого языка – его употребление закрепилось и сегодня является в корпусной лингвистике устоявшимся. В этом месте как раз полезно сформулировать, что мы понимаем под корпусной лингвистикой, чтобы перейти к очерку краткой истории этой дисциплины.

В узком смысле корпусную лингвистику обычно понимают практически, то есть как технологию создания и использования электронных корпусов текстов. В таком понимании это сугубо прикладная дисциплина, отвечающая не столько за получение и представление "знаний", сколько за разработку "умений" (того, что обычно называют английским термином "know-how"). Однако существует и более широкое понимание задач корпусной лингвистики, при котором она соотносится с методологией и результатами исследования языка с помощью корпусных инструментов. При таком понимании из прикладной области корпусная лингвистика становится особым направлением теоретического языкознания со своим представлением о сущности языка и со своими методологическими постулатами. На первый взгляд такие масштабные претензии корпусной лингвистики могут показаться (и, действительно, казались многим лингвистам в начале корпусной эпохи) не вполне мотивированными: ведь корпус – это как будто бы не более чем инструмент, который никак не влияет на качество полученных данных. Доступность этих данных, конечно, становится большей, но разве это ведёт к каким-то более глубоким изменениям лингвистики в целом? Разве с помощью корпуса лингвисты не продолжают просто делать всё то же, что и раньше, только быстрее и эффективнее? Если бы это было так, то, действительно, в термине "корпусная лингвистикам было бы не очень много смысла – мы же не говорим, например, о "табличной лингвистике", несмотря на то что лингвисты чрезвычайно успешно пользуются таблицами для представления данных о лексике и грамматике языка. Но выбор таблицы как способа представления данных, в общем и целом, не меняет сам подход к этим данным, и данные не становятся

[2] В реальном корпусе (например, в Национальном корпусе русского языка, о котором подробнее ещё будет сказано) такой запрос выполняется за доли секунды и позволяет убедиться в том, что у Тургенева в текстах данного периода розы бывают белые, красные и лиловые. Конечно, эвристическая ценность этого факта может показаться не вполне очевидной, но в данном случае он призван лишь проиллюстрировать бесконечно разнообразные возможности корпусного поиска и лёгкость его использования. О некоторых гораздо более серьёзных результатах, которые может давать корпусное исследование языка, пойдёт речь ниже.

от этого другими[3]. В случае с использованием корпусов дело, однако, устроено принципиально иначе, хотя оппонентам корпусной лингвистики (а к ним в своё время принадлежал даже Ноам Хомский) первоначально это было не очевидно. Сегодня не подлежит сомнению, что корпус – это не просто новый инструмент, а фундамент новой теории языка: в начале XXI в. корпусные технологии и методы заставили лингвистов изменить очень многие привычные взгляды на природу языка и законы его устройства и изменения. Складывающиеся контуры новой теории языка (учитывающей данные корпусов и возможности корпусных методов) противостоят доминировавшим в середине XX в. взглядам на язык – это были структурные и ранние генеративные теории. В наибольшей степени корпусно-ориентированные теории языка близки тем направлениям теоретической лингвистики, которые называют себя функциональными (и среди них в особенности моделям "лингвистики узуса", или usage-based approaches [7, 8]). Сближение корпусной и функциональной лингвистики узуса – скорее, встречное движение: с одной стороны, широкое использование корпусов способствовало утверждению тех взглядов на теорию языка, которые независимо развивались создателями моделей узуса; с другой стороны, именно функциональные лингвисты наиболее интенсивно пользовались корпусами (и квантитативными моделями [9, 10]) в своей исследовательской практике.

Но такова корпусная лингвистика сегодня. В начале же её возникновения ничто как будто бы не предвещало столь успешной карьеры. Первые электронные корпуса языков начали появляться в 1960-е годы (для несовершенных ещё компьютеров того времени); их отличал небольшой объём и ограниченный функционал применения. Так, первый известный корпус, так называемый "брауновский" (созданный в 1961 г. в Брауновском университете США), насчитывал всего 1 млн словоупотреблений и состоял из 500 текстовых фрагментов равного объёма по 2 тыс. слов. Подобные корпуса обычно использовались для далёких от "мейнстримной" лингвистики того времени задач (например, с их помощью определялась относительная частота букв в английском тексте) и считались маргинальным вспомогательным инструментом для статистики, имеющим ценность только для решения прикладных задач. Структурная лингвистика XX в. вообще крайне скептически относилась к статистическим моделям (отдельные исключения, конечно, были, но для общего теоретического духа они были нехарактерны), считая, что для понимания языка они ничего дать не могут; тот "квантитативный поворот" [11], который стал визитной карточкой лингвистики XXI в., требовал совсем других инструментов – и других корпусов. Такие корпуса появились в 1990-е годы, в эпоху персональных компьютеров и Интернета, сразу способствовавшего возникновению огромного общедоступного массива оцифрованных текстов (исследователи языка сами по себе эту задачу никогда бы в исторически короткие сроки, конечно, не решили). Корпуса эпохи Интернета отличаются большим объёмом (от 100 млн словоупотреблений, как у BNC в начале 1990-х годов, до миллиардов, как у современных веб-корпусов), сложным и разнообразным функционалом и способностью представлять разные типы текстов (существуют письменные, устные и мультимедийные корпуса, параллельные многоязычные корпуса и другие разновидности). Все эти корпуса активно используются не только для исследования и документации языков (сегодня, наверное, ни одно масштабное исследование, особенно если речь идёт о крупных языках, уже не делается вне корпуса), но и в преподавании языка. Можно сказать, что современная лингвистика в огромной степени обязана своим успехам развитию современных компьютерных технологий, которые, придя в изучение языка, изменили лингвистику изнутри. Для многих лингвистов этот результат был неожиданным.

Стоит кратко коснуться и истории корпусной лингвистики в СССР и в России, где её развитие имело свои особенности и нюансы. Подробнее российская корпусная лингвистика (и её главный продукт – Национальный корпус русского языка) освещается в специальной статье Е.В. Рахилиной в настоящем выпуске "Вестника РАН", здесь же мы напомним лишь основные вехи.

На уровне идей отечественная наука до конца 1980-х годов шла вровень с веком и даже в чём-то опережала мировые разработки. В этом контексте особенно важна новаторская идея "машинного фонда русского языка", выдвинутая академиком А.П. Ершовым (1931–1988) в начале 1980-х годов [12, 13]. Машинный фонд, первоначально понимавшийся как максимально представительная оцифрованная коллекция русских текстов, призванная автоматизировать главным образом работы по лексикографии, довольно быстро в основных чертах оказался идентичен по замыслу большим корпусам, которые появятся в мире в течение следующего десятилетия. К сожалению, полноценной реализации этой идеи помешала как ранняя смерть А.П. Ершова (он прожил всего 57 лет), так и отсутствие у нас в стране

[3] Впрочем, даже и это утверждение верно лишь до известного предела: иногда выбор подходящего табличного формата представления данных (прежде всего в случае флективных словоизменительных парадигм) позволяет увидеть такие особенности, которые остаются незамеченными при других подходах. В особенности это существенно для описания грамматического синкретизма (например, [5]); применительно к русской падежной системе этот парадокс в заострённом виде обсуждается в работе [6]. В лингвистике вообще связь инструмента описания с результатами описания нередко оказывается сильнее, чем во многих других науках.

в то время подходящих технологий для оцифровки текстов в больших масштабах. “Второй раунд” реализации идей корпусной лингвистики применительно к русскому языку наступил только в начале 2000-х годов; его воплощением и стал проект Национального корпуса русского языка, который возник в ходе тесного сотрудничества лингвистов (главным образом выпускников кафедры теоретической и прикладной лингвистики МГУ им. М.В. Ломоносова) и программистов компании “Яндекс” под руководством И.В. Сегаловича: обе группы исследователей понимали необходимость создания корпуса и нуждались друг в друге. Аналогичные попытки делались в то время и другими коллективами, но именно данному проекту сопутствовала удача – в частности, на самых ранних этапах он был поддержан не только компанией “Яндекс”, но и Российской академией наук и руководством Института русского языка им. В.В. Виноградова РАН (где выполнялся и проект Машинного фонда, материалы которого удалось частично интегрировать в Национальный корпус русского языка). В результате Национальный корпус русского языка оказалось возможным создать в очень короткие для проектов такого рода сроки: первые рабочие семинары с обсуждением конкретных шагов начались в 2001 г., а уже 9 апреля 2004 г. корпус был открыт для свободного доступа в Интернете. До сих пор судьба его складывалась счастливо (несмотря на неизбежные трудности и риски, не раз возникавшие на его пути): за 20 лет существования он стал одним из наиболее востребованных рабочих инструментов для русистов во всём мире, и сегодня, наверное, нет таких исследований русского языка, которые выполнялись бы без обращения к корпусу.

Сравнительно позднее вхождение русского языка в клуб языков, обладающих национальными корпусами, имело и свои преимущества: создатели смогли учесть мировой опыт и выбрать оптимальную траекторию развития, не повторяя ошибок прежних разработчиков и улучшая то, что можно было улучшить. Можно сказать, что, по сравнению с поколением корпусов 1990-х годов, Национальный корпус русского языка воплощает новую концепцию: фактически это первый корпус, созданный лингвистами для “обычных” пользователей, не обладающих специальной компьютерной подготовкой; он предназначен в первую очередь для задач научного исследования языка во всех его аспектах, для чего привычный корпусный функционал был значительно обновлён и усовершенствован; собственно, этот процесс продолжается и в настоящее время.

Итак, современная корпусная лингвистика – это новое направление теоретических исследований естественного языка, предоставляющая в распоряжение специалистов принципиально новые данные. Связано это в первую очередь с объёмом корпусного материала: он на много порядков превосходит ту информацию, которую человек может получить и обработать, не прибегая к услугам современных технологий. Поэтому в докорпусную эпоху многие вопросы, которые сейчас стоят на повестке дня, не могли быть не только разрешены, но даже и поставлены. Лингвистика XX в. имела отчётливо субъективный характер: в отсутствие доступа к “большим данным” исследователь вынужден был обращаться прежде всего к интроспекции[4], то есть к собственному представлению о том, как должны выглядеть “грамматически правильные” высказывания. В результате теории языка XX в. были в основном теориями того, чего в языке не может быть (это называлось “исследованием ограничений” и считалось крайне важной чертой “настоящей” научной теории): теоретическая лингвистика постепенно превратилась в науку о несуществующем.

Появление корпусов изменило повестку прежде всего в этом аспекте: корпус помог вернуть лингвистам их подлинный (и, может быть, самый важный) объект, то есть тексты на естественных языках без всяких ограничений на их объём. Это позволило постепенно приблизить лингвистику к идеалу эмпирических естественных наук (каковой она почти всю свою историю, вообще говоря, стремилась стать): лингвисты смогли задавать вопросы о том, что в языке реально существует (и количественно оценивать частотность таких явлений). Приоритетным объектом теорий узуса являются не “ограничения”, а частотные явления: язык – это механизм, который позволяет порождать прежде всего частотные элементы корпуса, а следовательно, они и должны изучаться в первую очередь. Более того, такие утверждения можно надёжным образом проверить, не обращаясь к субъективной интроспекции. Не менее важной лингвистической информацией, предоставляемой корпусом, оказывается информация о контекстном окружении языковых единиц, или их “сочетаемости”: так, задав соответствующие запросы, можно почти мгновенно узнать, какие объекты в текстах на русском языке бывают *круглыми*, какие процессы могут происходить *быстро* или *медленно*, какие ситуации могут иметь место *с... по...* определённый временной период (и какой именно) и т.д. В традиционных грамматиках и словарях такая информация отражается, как правило, очень избирательно и неполно.

Наконец, корпус даёт возможность осуществлять точные и эффективные наблюдения за динамикой

[4] Некоторым исключением являлись те области лингвистики, которые имели дело с данными мёртвых языков (такие, например, как классическая филология), где интроспекция была по понятным причинам недоступна: фактически в этих областях многие идеи и методы корпусной лингвистики использовались, так сказать, *avant la lettre*. Но не эти области определяли основную теоретическую повестку лингвистики XX в.; кроме того, работа с корпусами мёртвых языков в отсутствие современных технологий также имела свои ограничения.

языковых изменений, что делает современную теорию языка прежде всего диахронической теорией: изменение во времени – существенное и неотъемлемое свойство любого человеческого языка. Между тем, теории XX в. этот аспект традиционно игнорировали в пользу "строго синхронного" описания; сегодня ясно, что такой методологический ригоризм ничем не оправдан, более того, "строго синхронный" срез языка является несуществующим конструктом. Но для построения полноценной диахронической теории ранее не хватало нужных инструментов – именно их мы сегодня находим в исторических корпусах языков. Особенно хорошо корпус позволяет прослеживать так называемые микродиахронические изменения, затрагивающие малозаметные для обычного носителя сдвиги в значении и сочетаемости с шагом в несколько десятилетий; это возможно благодарю тому, что тексты, входящие в корпус, датированы (и, как правило, достаточно точно). Особенно эффектно, когда такие изменения возникают буквально на наших глазах, и корпус позволяет присутствовать непосредственно при их зарождении.

Одним примером такого явления мы бы и хотели завершить настоящие заметки. В русском языке имеется частотный глагол *"уметь"*, относящийся к базовой лексике и, казалось бы, не таящий особых сюрпризов. Стандартная конструкция, в которой этот глагол участвует, выглядит так: *кто-то* [субъект, обычно личный, выраженный номинативом] *умеет делать что-то* [прямой объект, выраженный инфинитивом со своими дополнениями]. Конструкция описывает способности субъекта (обычно приобретённые в результате обучения); многочисленные корпусные примеры её употребления выглядят, например, так:

(1) «*Я* ***умею*** *играть "Moon River" на пианино*» [А. Геласимов. Нежный возраст. 2001].

(2) "Лисевицкий не ***умел*** *играть в покер*, и его дружно учили" [В. Лихоносов. Ненаписанные воспоминания. 1983].

(3) "...в чужих людях без средств к существованию не прожить, а Ольга не ***умела*** даже *помыть посуду*" [В. Пьецух. Шкаф. 1997].

Иногда глагол *"играть"* (и только он!) при названии игры может опускаться:

(4) "Лаврецкий подсел к ней и стал глядеть ей в карты. – Вы разве ***умеете в пикет***? – спросила она его с какою-то скрытой досадой" [И. Тургенев. Дворянское гнездо. 1859].

(5) "Ты же в теннис не играешь, – смеялась она, – а я не ***умею в гольф***" [Н. Климонтович. Парадокс о европейце. 2013].

То, что мы наблюдаем в примерах (4) и (5), обычно называется "эллиптическими конструкциями"; опущенный элемент в них однозначно восстанавливается на уровне смысла: *не умею в гольф* = не умею **играть** в гольф. Внешне, однако, эллиптические конструкции типа *"уметь в гольф"* выглядят как "обычные" непереходные конструкции с предлогом (наподобие таких, как *"смотреть в окно"* или *"отойти в сторону"*). Но это сходство поверхностное: в случаях типа *"смотреть в окно"* предложная группа действительно синтаксически зависит от глагола (*смотреть – во что?*), тогда как в случаях типа *"уметь в гольф"* предложная группа синтаксически зависит не от глагола *"уметь"*, а от опущенного (но подразумеваемого) инфинитива *"играть"*. Тем не менее само наличие такого сходства, как оказалось, запустило в языке некие сложные механизмы синтаксических изменений, при которых интерпретация эллиптической конструкции подверглась влиянию более распространённой непереходной конструкции с предлогом (такой процесс в лингвистике обычно называют "ре-анализом"). И вот такой процесс происходит в русском языке, можно сказать, буквально на наших глазах: возникает и широко распространяется новая конструкция, в которой глагол *"уметь"* регулярно сочетается с предложной группой, уже не предполагая никакого эллипсиса. Сам глагол *"уметь"* при этом приобретает новое значение, не очень резко, но достаточно заметно отличное от имевшихся у него ранее. В этой новой нестандартной конструкции *"*(*кто-то*) ***умеет во что-то"*** значение глагола приблизительно сводится к смыслу "владеет, разбирается, способен применить"; например, *"я не умею в сарказм"* означает что-то вроде "я не владею приёмами сарказма", *"я не умею в интегралы"* употребляется, когда хотят сказать "я не разбираюсь в интегралах и не умею ими пользоваться".

Перед нами – достаточно типичный процесс языковых изменений; подобные им постоянно происходят во всех естественных языках на протяжении всей их истории. Примечательным в данном случае является то, что мы с помощью корпуса можем точно засвидетельствовать начало этих изменений и понять, как они происходят.

В нашем случае источником инновации, как и во многих других, является молодёжный сленг – младшее поколение носителей языка вообще, как известно, является наиболее мощным триггером языковых изменений, которые впоследствии закрепляются в речи тех же носителей в более старшем возрасте и из инновации становятся нормой. В современном русском языке эта стадия для указанной конструкции ещё не наступила, поэтому она пока сохраняет явно выраженный сленговый характер и носителями более старших поколений воспринимается как "неправильная". Но мы, лингвисты, хорошо знаем, что язык – явление природы, и в нём не бывает "правильных" и "неправильных" компонентов: есть

лишь непрерывно меняющаяся реальность, и совершенно естественно, что инновационные явления частью общества оцениваются негативно (нам почти всегда кажется, что тот, кто говорит не так как мы, говорит "плохо"). С точки зрения научного изучения языка, все факты языка одинаково ценны; более того, любые изменения языка подчиняются определённым законам и не могут происходить произвольно. Корпус, как мощное увеличительное стекло, даёт нам возможность рассмотреть эти процессы много более детально, чем могли это сделать лингвисты прошлого.

Согласно корпусным данным, первые надёжно документированные примеры новой конструкции *"уметь в"* относятся только к началу 2010-х годов, то есть являются совсем недавними. Их источник – в основном, социальные сети: хорошо известно, что язык Интернета очень полно и гибко отражает языковые инновации, ведь в соцсетях говорящие обычно фиксируют на письме свою спонтанную речь, не очень задумываясь о литературных стандартах – а это и есть самое ценное для лингвиста. Поскольку в Национальный корпус русского языка в настоящее время включён достаточно представительных подкорпус языка русского сегмента Интернета, появление таких конструкций и динамика их развития фиксируется корпусными методами достаточно надёжно. Вот некоторые реальные примеры из корпуса (в скобках даётся точная дата фиксации реплики; возможность в большинстве случаев установить такую дату является ещё одним преимуществом современных соцсетей):

(6) "Прекрасные стрелки. Выносливы, храбры. Умеют действовать группами, в одиночку, мотивированы и коварны. В бою действуют грамотно, ***умеют в тактику и стратегию***. Не пугаются техники и авиаударов, эффективно действуют из засад" [16.07.2014].

(7) "Как мало женщин ***умеют в комедию***. Поэтому сегодня мы поздравляем ту, у которой это получается лучше всех. Очаровательная и харизматичная Вупи Голдберг, подарившая нам множество улыбок в детстве, сегодня отмечает своё 59-летие" [13.11.2014].

(8) "Ребята, кто-нибудь ***умеет в мобильные сайты***? Как написать код, который... Там строчки три примерно, мы не можем дружно вспомнить, как это. Буду крайне благодарен за помощь" [07.05.2015].

(9) "В разные моменты своей жизни я либо ***умею в тайм-менеджмент*** очень хорошо, либо не умею вообще. Середины практически нет" [21.12.2015].

(10) "К. – актёр?! А вы ***умеете в чёрный*** юмор" [29.06.2022].

Как можно видеть, данная конструкция прочно укоренилась среди носителей младшего поколения и приближается по своему статусу к нейтральной (проникая уже и в некоторые электронные СМИ); у неё хорошие перспективы в скором времени (то есть через одно-два поколения говорящих) стать нормативной и обогатить базовый инвентарь русских модальных конструкций – хотя, конечно, в истории языковых изменений для нас и сегодня много неясного, и однозначно предсказать её судьбу мы не можем.

* * *

Современные электронные корпуса поддерживают все виды лингвистических исследований и в настоящее время никакое серьёзное исследование языка без них невозможно. Существующие компьютерные технологии позволяют шире использовать очень нужные для теоретической лингвистики квантитативные методы и облегчают визуализацию данных (в современных корпусах этот приём активно используется). Но не следует забывать, что сегодня корпуса не являются достоянием одних только лингвистов-теоретиков – они активно внедряются в преподавание родного языка и технологии обучения иностранным языкам. Фактически они, как уже было сказано, формируют новую идеологию современной лингвистики: описание языка теперь понимается как описание *на основе определённого корпуса*. Такое описание нацелено прежде всего на частотные явления, и оно становится проверяемым, то есть более точным и объективным. Таким образом, приход корпусов в лингвистику ("корпусная революция") изменил не только инструменты получения языковых данных, но и теоретические установки исследователей.

В своё время академик Ю.Д. Апресян сформулировал важнейший принцип описания языка, который он назвал "интегральным" (первое подробное изложение этой идеи дано в статье [14]; в дальнейшем она получила развитие во многих работах как самого Апресяна, так и других представителей Московской семантической школы). Суть этого принципа в том, что лингвист должен стремиться строить описание языка на основе двух тесно связанных модулей: к ним относится *грамматика* (понимаемая в общих чертах как система правил, по которым строятся тексты) и *словарь* (те элементы, из которых строятся тексты по правилам грамматики). Интегральное описание требует, чтобы словарь составлялся с учётом грамматики, а грамматика была ориентирована на конкретный словарь[5], тогда информация о языке будет полной и такое описание обеспечит решение многих важных теоретических и прикладных задач, тогда как в традиционной лингвистике это обычно изолированные друг от друга

[5] "Интегральным, или единым, мы будем называть такое лингвистическое описание, в котором грамматика и словарь согласованы друг с другом по типам помещаемой в них информации и по формальным языкам её записи" [14, с. 57].

области. Сегодня мы говорим, что в интегральное описание языка обязательно должен входить третий компонент, и это, как нетрудно догадаться, *корпус*. В рамках интегрального описания корпус понимается как полная совокупность тех текстов, на которых проверяются словарные и грамматические модули: всякое утверждение, сделанное в словаре и грамматике, должно подтверждаться примерами из соответствующего корпуса (то есть перед нами на самом деле корпусная грамматика и корпусный словарь, интегрированные в единое целое). Это ещё один очень важный шаг к повышению научной достоверности и объективности лингвистических описаний и лингвистических теорий.

В ближайшей – или, может быть, более отдалённой – перспективе для всех языков мира (и, безусловно, для всех языков России) должны существовать интегрированные словари, грамматики и корпуса. Как кажется, мировое лингвистическое сообщество понимает эту задачу и движется в направлении её реализации.

ЛИТЕРАТУРА

1. *Плунгян В.А.* Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении. 2008. № 2 (16). С. 7–20.

   Plungian V.A. Corpus as a tool and as an ideology: on some lessons of modern corpus linguistics // Russian language in scientific coverage. 2008, no. 2 (16), pp. 7–20. (In Russ.)

2. *Плунгян В.А.* О перспективах современной корпусной лингвистики // Труды Отделения историко-филологических наук РАН / Под. ред. В.А. Тишкова. М.: Наука, 2016. С. 128–132.

   Plungian V.A. On the prospects of modern corpus linguistics // Proceedings of the Department of Historical and Philological Sciences of the Russian Academy of Sciences / Ed. by V.A. Tishkov. Moscow: Nauka, 2016. Pp. 128–132. (In Russ.)

3. *Аброскин А.А.* Поиск по корпусу: проблемы и методы их решения // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы / Отв. ред. В.А. Плунгян. СПб.: Нестор-История, 2009. С. 277–282.

   Abroskin A.A. Corpus search: problems and methods of their solution // National Corpus of the Russian language: 2006–2008. New results and prospects / Ed. by V.A. Plungian. St. Petersburg: Nestor-Istoriya, 2009. Pp. 277–282. (In Russ.)

4. *Čermák F.* Today's corpus linguistics: Some open questions // International journal of corpus linguistics. 2002, vol. 7, pp. 265–282.

5. *McCreight K., Chvany C.V.* Geometric representation of paradigms in a modular theory of grammar // Paradigms: The economy of inflection. Berlin: Mouton de Gruyter, 1991. Pp. 91–112.

6. *Плунгян В.А.* Геометрия русского словоизменения: о традиционных и нетрадиционных таблицах склонения // Зборник Матице Српске за славистику. 2021. Т. 100. С. 187–203.

   Plungian V.A. The geometry of Russian inflection: on traditional and non-traditional declension tables // Zbornik Matice Srpske za Slavistiku. 2021, vol. 100, pp. 187–203. (In Russ.)

7. *Barlow M.* Ten lectures on corpora and cognitive linguistics. Leiden: Brill, 2023.

8. *Hollmann W.B.* Generative vs. usage-based approaches to language // Introducing linguistics, ch. 25. London: Routledge, 2022.

9. *MacWhinney B.* Emergentist approaches to language // Frequency and the emergence of linguistic structure (Typological studies in language 45). Amsterdam: John Benjamins, 2001. Pp. 449–470.

10. *Bybee J.* Language, usage and cognition. Cambridge: Cambridge University Press, 2010.

11. *Kortmann B.* Reflecting on the quantitative turn in linguistics // Linguistics. 2021, vol. 59, pp. 1207–1226.

12. *Ершов А.П.* Машинный фонд русского языка (внешняя постановка вопроса) // Вопросы языкознания. 1985. № 2. С. 51–54.

    Ershov A.P. The machine fund of the Russian language (external formulation of the question) // Voprosy Jazykoznanija. 1985, no. 2, pp. 51–54. (In Russ.)

13. *Андрющенко В.М.* Концепция и архитектура Машинного фонда русского языка. М.: Наука, 1989.

    Andryushchenko V.M. The concept and architecture of the Machine Fund of the Russian language. Moscow: Nauka, 1989. (In Russ.)

14. *Апресян Ю.Д.* Интегральное описание языка и толковый словарь // Вопросы языкознания. 1986. № 2. С. 57–70.

    Apresyan Yu.D. Integral description of language and explanatory dictionary // Voprosy Jazykoznanija. 1986, no. 2, pp. 57–70. (In Russ.)

CORPUS LINGUISTICS NOWADAYS

V.A. Plungian [a,b,*]

[a] *Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia*
[b] *Institute of Linguistics of the Russian Academy of Sciences, Moscow, Russia*
**E-mail: plungian@iling-ran.ru*

The article proposes a general presentation of corpus linguistics, its history, its methodology and its influence on current views on how a language must be studied – the process which is usually referred to as “corpus revolution”.

Keywords: corpus linguistics, linguistic theory, linguistic methodology, functional linguistics.