НАЦИОНАЛЬНЫЙ КОРПУС русского языка

Александр Молдован,
директор Института русского языка им В.В. Виноградова Российской академии наук, член-корреспондент РАН, доктор филологических наук

Грамматика и словарь — с давних пор традиционные формы представления языка. Уменьшенные копии грамматик и словарей, дополненные орфографическими правилами, становились школьными учебниками. В отличие от уроков химии или физики, на которых школьники на собственном опыте могут убедиться в действии тех или иных законов, на уроках русского языка они ограничены лишь задачей освоения литературных норм, извлекаемых исключительно из учебника. До тех пор, пока описание языка строилось на основе индивидуального опыта и интуиции учёного-языковеда, ничего другого в помощь школе лингвистика не могла предложить. Положение изменилось благодаря компьютерам, когда появилась третья форма представления языка — его корпус.

Национальным корпусом русского языка называется общедоступная справочно-информационная система по русскому языку, которая была создана коллективом специалистов Института русского языка им В.В. Виноградова РАН и других академических лингвистических институтов и университетов России и уже несколько лет действует в Интернете по адресу www.ruscorpora.ru при поддержке компании «Яндекс».

Что же такое корпус вообще и Национальный корпус в частности? Какие задачи он решает и какой может быть его роль в образовании?

Корпус языка — это собрание текстов в электронном виде, но не обычное, а с приписанной к каждому тексту и каждому слову научной информацией о характеристиках и свойствах данного текста и данного слова. Этот аппарат называется «разметкой». Корпус тем лучше, чем полнее и совершеннее его разметка. Существует новая отрасль языкознания — корпусная лингвистика — наука о том, как сделать хорошую разметку корпуса. Зачем она нужна?

Хорошая разметка позволяет быстро и эффективно найти в корпусе те слова, формы и конструкции, которые нужны исследователю. Для этого программа поиска должна «понимать» как минимум то, какие формы в тексте относятся к одному и тому же слову. Например, если нас интересует глагол *вести*, то программа должна нам выдать контексты не только с формой инфинитива *вести*, но и со всеми

остальными формами (*веду*, *ведёшь*, *ведущий*, *ведомый* и т.д.), при этом она не должна выдавать, например, *ведаю*, *ведунья* или *вести «новости»*. Обычные поисковые системы этого делать не могут, так как для этого необходимо, чтобы компьютер «понимал» грамматическую структуру языка.

Тем более это понимание необходимо, если мы хотим искать не конкретные слова, а отвлечённые грамматические формы. Представьте, что нам нужно найти в достаточно длинном тексте все слова в дательном падеже единственного числа. Текстовый редактор такие задачи решать не умеет. Для того чтобы грамматические формы можно было автоматически найти в тексте, эту информацию необходимо предварительно в него ввести. Иначе искать придётся только вручную, а эта процедура долгая и трудоёмкая. Поэтому с самого начала работ над Корпусом лексико-грамматическая разметка (приписывание каждой словоформе информации о лексемной принадлежности и об инвентаре морфологических признаков) была главным направлением работы. Для большинства слов эта задача была решена автоматически — с помощью специально разработанных программных средств морфологического анализа. В тех случаях, когда алгоритмы разметки не давали однозначного ответа, вступала в действие технологическая цепочка ручного приписывания каждой неоднозначной словоформе (омографу) правильного разбора.

Кроме того, в семантическом разделе Корпуса осуществлена частичная разметка слов по семантическим признакам. Семантический поиск осуществляется по классам слов: наименования частей тела, имена родства, глаголы движения, оценочная лексика, слова со значением уменьшительности и т.п.

Наконец, разметка включает метатекстовую информацию. Пользователь может быть заинтересован в том, чтобы ограничить поисковую выдачу по самым разным параметрам: например, чтобы поиск заданной им комбинации грамматических и лексических признаков осуществлялся только в мемуарах или только в записях устной речи, игнорируя, например, длинный ряд однородных примеров из газетного текста (на сайте это ограничение задаётся в разделе «Мой корпус». Метатекстовая информация включает параметры: автор (имя, пол, возраст), название, дата создания, объём (в словах) текста; жанр, тип текста (рассказ, роман и т.п.), место и время описываемых событий; для нехудожественных текстов — сведения о функциональной сфере, типе и тематике текста и т.п. По всем этим метатекстовым параметрам возможны поиск и создание пользовательского подкорпуса. Например, вас интересует, изменилась ли на протяжении того или иного отрезка времени частотность употребления в русском языке слова *порядочность* (интуиция подсказывает, что за последние годы она изменилась). Выбираем хронологические границы «до 1990 г.» и проводим поиск. Обнаруживается, что из 9590 документов до 1990 г. слово *порядочность* встречается в 182-х. Это составляет приблизительно 1,89%. Потом смотрим документы «после 1990 г.». Из 34 390 документов последнего времени слово *порядочность* встречается в 226, что составляет 0,66%. Так выясняется, что это слово за последние 18 лет стало употребляться в три раза реже, чем прежде.

Электронное собрание текстов

Итак, корпус — это электронное собрание текстов, размеченное таким образом, чтобы в нём можно было быстро найти слова и конструкции с заданными грамматическими и другими свойствами. *Национальным* корпусом, по сложившейся традиции, называют самый большой и представительный корпус, характеризующий язык данной страны в целом. Сегодня большинство крупных языков мира уже имеет свои национальные корпуса. Общепризнанным образцом является, в частности, Британский национальный корпус (BNC), своими национальными корпусами располагают Америка, Германия, Италия, Испания, Венгрия, Литва, Эстония, Ирландия и другие (в том числе славянские) страны. Наш национальный корпус сопоставим с этими корпусами по

объёму текстов (его объём сегодня составляет более 140 млн словоупотреблений, и он продолжает пополняться), но при этом наш корпус значительно превосходит большинство зарубежных по детальности разметки и, следовательно, возможностям поиска. В частности, в Национальном корпусе русского языка есть возможность точного грамматического поиска по очень большому массиву словоупотреблений. Кроме того, осуществляется семантический поиск по классам слов — можно выбрать, допустим, названия частей тела, термины родства и т.п. Уникальной особенностью нашего корпуса является то, что в нём возможен сложный поиск, т.е. поиск языковых конструкций длиной до 10 слов с заданной комбинацией характеристик для каждого компонента такой конструкции и заданным расстоянием между словами.

Помимо того, что корпус должен быть большим, он — и это даже важнее — должен быть представительным, *репрезентативным*. Иначе говоря, он должен содержать все типы текстов, представленные в данном языке в данный исторический период, и при этом содержать их в правильной пропорции, должен быть *сбалансированным*.

Именно поэтому Национальный корпус русского языка не ограничивается только произведениями художественной литературы XIX и XX века, сколь бы важны они ни были для изучения русского языка. Он содержит и газетные и журнальные статьи разной тематики (от общественно-политических до, например, спортивных), и специальные тексты (научные, научно-популярные и учебные по разным отраслям знания), и рекламу, и частную переписку, и дневники. Словом, в Корпус попадают образцы практически любого существующего в русском языке письменного дискурса — от статьи современного музыкального критика до инструкции по уходу за кактусами, от классической литературы до справочника по физике.

Тексты представлены в определённой пропорции, отражающей их долю в общем массиве текстов данной эпохи. Так, доля художественных текстов (включая драматургию и мемуары) будет составлять около 40% (в настоящее время она несколько выше, что связано с особенностями процесса пополнения Корпуса новыми текстами). Интересно, что в корпусах европейских языков эта доля существенно ниже — как правило, она не превышает 20%. Так что Национальный корпус русского языка всё равно остаётся одним из самых «литературоцентричных». Но нужно понимать, что писатели — во всяком случае те, что считаются наиболее интересными — скорее выступают как экспериментаторы, языковые эквилибристы, всевозможными способами нарушающие языковые нормы. Конечно, нарушать нормы лингвисту не менее интересно, чем их соблюдать, но в первую очередь должен быть представлен нормативный языковой «фон» данной эпохи. Очень хороший баланс представляют в этом отношении мемуары: это тексты с элементами художественной организации, написанные, как правило, людьми незаурядными, хорошими рассказчиками, но при этом тексты, гораздо более укоренённые в языковой стихии повседневности, чем современная художественная литература.

Всё это богатство текстов, собранное вместе и отражающее современный русский язык во всех его аспектах, стилях, жанрах и формах существования, находится в свободном доступе в Интернете, что позволяет любому человеку искать любые слова, формы или сочетания слов в определённой грамматической форме или выяснять, каким материалом представлена та или иная грамматическая категория. Кроме того, можно самостоятельно выбирать нужное подмножество текстов и искать, например, тексты определённого автора, определённого периода, определённого жанра и т.п. в любых комбинациях.

Создание национального корпуса — культурно значимое событие, вопрос престижа языка. Но дело не только в престиже.

Если в целом оценивать значение Корпуса для русистики, то без преувеличения

можно сказать, что с появлением Национального корпуса мы впервые получаем материальную базу для объективных и достоверных суждений о современном состоянии и путях развития русского языка. Это значение покрывает всё остальное, потому что Корпус тем самым становится новой и важнейшей частью материального представления русского языка. Более того, согласно сегодняшним мировым стандартам, только корпус может служить основой для словарей, грамматик и учебников нового поколения. Корпус позволяет обнаружить в языке такие закономерности, о существовании которых не было известно или они не могли быть строго обоснованы. Тем самым он позволяет ставить новые проблемы, которых лингвистика прошлого почти не касалась.

Эволюция значения слова

Одно из таких новых направлений — это наблюдение за *динамикой* развития языка, своего рода «микроисторическая» лингвистика, когда в центре внимания находятся не глобальные изменения в истории языка, а изменения, происходящие буквально на наших глазах, занимающие одно-два десятилетия (для истории языка это чрезвычайно маленький срок). Поскольку тексты, входящие в Корпус, датированы, нетрудно проследить за хронологией постепенных языковых изменений — за появлением или постепенным угасанием определённых слов и конструкций, за изменением ими своих значений и т.п.

Примером использования Корпуса для изучения эволюции значения слова может быть поиск по слову *истребитель*. Это слово представлено в Корпусе 1832 примерами. Выявить их и расположить в хронологическом порядке вручную было бы практически невозможно. Корпус позволяет это сделать автоматически — выбором опции «Сортировать по дате создания». Упорядочив таким образом выдачу примеров от Карамзина до наших дней, можно обнаружить, что в XIX — начале XX века это слово употреблялось только в значении «тот, кто истребляет» (Н.М. Карамзин: «От Клина до Городни и далее истребители шли с обнажёнными мечами, обагряя их кровию бедных жителей»; А.М. Горький: «Все мои симпатии были и будут на стороне несчастных армян, а не их зверских истребителей»).

Как название механизма, это слово впервые появляется только в 1923 году применительно к морскому судну: «А, это моторный катер, а может, и «истребитель» (И. Шмелёв. «Солнце мёртвых»). Судя по наличию кавычек, это слово тогда ещё было не освоено. В 1928 г. такое название судна употребляется уже без кавычек — у Андрея Платонова в «Сокровенном человеке» («Меж тем в порту появился маленький истребитель «Звезда»). Корабли-истребители фигурируют и в написанном в 1930-е годы романе А.С. Новикова-Прибоя «Цусима».

Первое упоминание истребителя как самолёта фиксируется только в знаменитом пассаже из «Мастера и Маргариты» М. Булгакова, написанном в 1929–1940 гг.: «Истребитель? Кто и в какой истребитель пустит Степу без сапог? Да и в сапогах в истребитель его не пустят!».

В период Великой Отечественной войны это значение становится повседневным: примеры на него встречаются в дневниках Всеволода Иванова, записных книжках А. Платонова, романе В. Каверина «Два капитана», у А. Фадеева, А. Бека, Э. Казакевича и многих других. Тогда же у этого слова появляется и значение «лётчик истребительной авиации». В вышедшем на экран в 1941 г. фильме «Валерий Чкалов» (в Корпусе представлены не только письменные тексты, но и сценарии фильмов) есть такие слова: «Я всем своим нутром, всеми печенками истребитель, боец!». В 1950–1990-е годы это значение встречается уже всегда в сочетании со словом лётчик: лётчик-истребитель («Красная звезда» за 1956 г., «Жизнь и судьба» Василия Гроссмана, 1960 г., «Тяжёлый песок» А. Рыбакова, 1974 г., воспоминания А.И. Микояна, 1971–1974 гг. и др.).

Эволюция грамматических форм

Самое простое, что сейчас можно узнать из корпуса о русском языке — это выяснить количественное соотношение употребления в русском языке различных частей речи. Обнаруживается, что «удельный вес» существительных в русских текстах составляет приблизительно 28,5%, прилагательных — 8,5%, глаголов в русских текстах — чуть более 17%, а наречий — меньше 4%. Предлоги занимают в текстах более 10%, союзы — почти 8%, частицы — около 5% и т.д. Раньше эта статистика была неизвестна.

Создание корпуса по-новому ставит одну из актуальных задач современной русистики — составление нового поколения академических словарей и нового описания грамматики русского языка[1]. Грамматические описания нового поколения должны будут не только учитывать изменения в русском языке за последние тридцать лет, не только использовать новые теоретические достижения лингвистической мысли, но и содержать новые обобщения, отражающие всесторонний анализ материалов Корпуса.

Для полного и адекватного представления о том, что происходит в современном русском языке, необходимо ещё в большей мере расширить рамки Корпуса. Поэтому помимо текстов, представляющих современный русский литературный язык (с начала XIX века), в Национальный корпус русского языка включаются также и тексты, представляющие древнерусский язык, и нелитературные формы современного русского языка: разговорную, просторечную, диалектную. Почему образцы устной речи так важно иметь в Корпусе? Люди пишут не так, как говорят; в особенности это различие ощутимо для языков с давней письменной традицией, за время существования которой нормы письменной и устной речи успевают разойтись достаточно сильно. Если мы хотим выявить наиболее динамичные структуры живого русского языка и хотя бы отчасти заглянуть в будущее русского языка, мы должны обратиться к стихии устной речи, не скованной традицией и нормой. Многое в устной речи поражает — но, с другой стороны, многие конструкции, существующие в современной устной стихии, неожиданно обнаруживаются в документах времён царя Алексея Михайловича и даже в новгородских берестяных грамотах XII—XIV вв. Отдельная проблема — включение в Корпус образцов не общерусского языка (пусть и в его разговорном варианте), а настоящей диалектной речи, которая также включается в Корпус.

В систему Национального корпуса русского языка входят несколько экспериментальных корпусов сравнительно небольшого объёма. Помимо диалектного, это параллельный русско-английский и англо-русский корпус (2 млн словоупотреблений), в котором текст на русском языке сопоставлен с переводом этого текста на английский язык и, наоборот, текст на английском языке сопоставлен с его переводом на русский язык. Кроме того, это корпус текстов XVIII века и, наконец, это очень интересный с технологической точки зрения экспериментальный корпус с синтаксической разметкой, созданный в Институте проблем передачи информации РАН.

Пользователи

Есть несколько больших групп специалистов, которым Корпус может быть полезен. Во-первых, это люди, в своей повседневной деятельности связанные со словом, прежде всего редакторы газет и журналов. Редакторам в своей практической деятельности постоянно приходится решать вопросы нормы: допустимо ли такое слово или конструкция? Кто, где, когда употребил впервые такой оборот? Для каких типов текста он наиболее характерен?

[1] На базе данных Национального корпуса русского языка Институтом русского языка им. В.В.Виноградова РАН уже сегодня созданы разнообразные новые словари современного русского языка: грамматический словарь новых слов русского языка, новый частотный словарь русской лексики, словарь русской идиоматики и др. (см.: http://dict.ruslang.ru). Эта работа продолжается.

Следующая по важности группа пользователей — те, кто так или иначе имеет дело с преподаванием современного русского языка, в том числе при дистанционном обучении. Многократно возрастает значение Корпуса при обращении к иностранной аудитории. Для людей, не владеющих русским языком в качестве родного (как преподавателей, так и учащихся), Корпус оказывается поистине незаменимым инструментом. Поэтому не случайна высокая популярность корпусов в иноязычной среде. И именно от зарубежных русистов (в особенности преподавателей русского языка) мы получаем очень заинтересованные отклики.

Для учителя школы, преподавателя вуза и студента Корпус ценен тем, что в нём можно быстро и легко найти целую серию примеров на трудное или редкое слово или грамматическую конструкцию и понять по контексту, как они употребляются — не только их значение, но и особенности их сочетаний, типы контекстов, социолингвистические характеристики и т.д.

Предположим, вас интересует слово «секьюрити»: что оно значит, какого оно рода и можно ли его склонять? Корпус выдаёт вам выборку из 62 текстов разного происхождения. Используя специальную возможность настройки, их можно расположить по хронологии. Это позволит обнаружить, что впервые интересующее нас слово появилось в книге Наталии Медведевой «Любовь с алкоголем» (1985 г.): «Даже люди из «секьюрити» сказали Наташке, что лучше бы им уйти». Кавычки, в которых появляется это слово, указывают на его новизну и неосвоенность в русской действительности и русском языке. При этом в Корпусе указано, что стиль данного текста — сниженный. Ещё в 1998 г. в мемуарах Алексея Козлова это слово встречается в кавычках: «У членов Политбюро и других партийных шишек, постоянно проезжавших по этой трассе, или у их «секьюрити», не дай бог, мог возникнуть вопрос: а почему это здесь толпится народ?». В то же время уже в 1996 году это слово фигурирует без кавычек в сочинениях Марины Вишневецкой и Николая Климонтовича. При этом у последнего оно дважды встречается в сочетании «служба секьюрити» («То, что Асанова готова сбагрить меня куда угодно — хоть в службу секьюрити, — было очевидно»; «Скорее всего, впрочем, её подставила служба секьюрити»). В опубликованных в 1997 г. мемуарах Вячеслава Фетисова «Овертайм» уже можно наблюдать, как русский язык трансформирует значение английского названия службы безопасности (security — «безопасность, защита»). Русское «секьюрити» становится названием не только службы безопасности, но и профессии охранника. Любопытно возникающее при этом колебание в выборе числа слов, согласуемых с этим несклоняемым существительным:

- *Секьюрити* сумасшедшие — стоят на каждом углу.
- Айрин встаёт посреди периода, зовёт *секьюрити*, тот к нам спускается, забирает у парня камеру, вынимает из неё плёнку, а его выгоняют с игры.

Эта числовая двойственность отражается и во множестве других текстов, в частности, у Виктора Баранца в мемуарах «Генштаб без тайн» (1999):

- Переводчик и *секьюрити* эскортировали меня к заветному туалетруму.
- Норвежец оказался сообразительным и учтивым человеком. Он что-то сказал *секьюрити*. Тот принёс мне сразу три бокала и что-то буркнул на своём языке.

Русский язык в Корпусе отражается таким, каков он на самом деле — во всём многообразии нормативных и ненормативных текстов. Это позволяет преподавателю показывать на уроках русского языка изменчивую среду, в которой формируются нормы литературного языка, демонстрировать роль образцовых текстов в этом процессе. Перед школьниками раскрывается реальное устройство русского языка и закономерности его функционирования. Изучаемый язык предстаёт не как нагромождение вокабул и грамматических правил, а как живой организм, открытый для обозрения и изучения.