Е. А. Гришина, В. А. Плунгян

ПЕРСПЕКТИВЫ РАЗВИТИЯ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА

Предыстория

Прежде чем говорить о перспективах развития Национального корпуса русского языка, необходимо напомнить некоторые внешние обстоятельства его создания.

Проект Национального корпуса русского языка возник и был реализован, может быть, в не совсем обычных обстоятельствах. Большинство участников этого проекта (за исключением С. А. Шарова) в начале работы не являлось специалистами по корпусной лингвистике в строгом смысле слова. Но все они были активно действующими профессиональными лингвистами и хорошо осознавали, насколько на современном этапе существования нашей науки такой инструмент, как корпус, необходим лингвисту в его повседневной деятельности — и морфологу, и синтаксисту, и лексикографу, и социолингвисту, и диалектологу.

К сожалению, в конце 1990-х гг., когда это осознание пришло, стало ясно также и то, что готовых электронных корпусов русского языка сколько-нибудь значительного объема, которыми можно было бы пользоваться, нет и в ближайшем будущем не предвидится. Выход был один — сделать собственный корпус, пригодный для решения тех задач, которые нам казались важными. Может быть, специалисты сделали бы эту работу лучше — но таковых не нашлось. Мы руководствовались известным принципом: «будем делать для себя — тогда и другим пригодится». Поэтому Национальный корпус русского языка делался быстро: мы не могли позволить себе роскошь обсуждать «теорию корпуса», основанную на бесконечной критике чужих результатов. Такое обсуждение, безусловно, приносит удовлетворение всем его участникам, но имеет только один изъян: каждый шаг в сторону еще большей теоретической безупречности не только не приближает к созданию корпуса, но почему-то ведет в противоположную сторону. После таких дис-

куссий корпуса не возникает. Возникают лишь бесчисленные проекты «самого лучшего в мире» корпуса. Негативный опыт такого рода имелся в достаточном количестве. Но нам не нужен был самый лучший в мире корпус. Нам не нужно было ощущение теоретического превосходства. Нам нужно было как можно скорее получить возможность искать надежные примеры в русских текстах.

Таким образом, Национальный корпус русского языка был сделан с ориентацией на сугубо прагматические соображения и в очень короткий срок. Разумеется, такой корпус не мог быть безупречным. Но мы стояли перед жестким выбором: или корпус будет иметь изъяны, или его не будет вовсе.

Теперь, однако, когда корпус существует и им можно пользоваться для решения уже очень многих и важных задач, можно — не выходя из той же прагматической логики — наметить некоторые наиболее актуальные направления дальнейшей работы. Это те наши пожелания, которые в силу технических или иных причин остались нереализованными на первом этапе, но которые продолжают оставаться для нас приоритетными. С их реализацией корпус, безусловно, станет лучше и удобнее, хотя это не означает, что он станет идеальным. Но в настоящих заметках мы хотели бы по понятным причинам ограничиться только теми направлениями будущего развития корпуса, которые непосредственно вытекают из опыта текущей работы. Более того, излагаемая ниже программа в той или иной степени уже частично выполняется: дополнительные усовершенствования, о которых пойдет речь, либо уже начали воплощаться в жизнь, либо активно обсуждались всеми участниками проекта на наших текущих семинарах. О задачах, относящихся к более дальней перспективе, речь практически идти не будет: это отдельная тема, которая требует иного подхода.

Ниже перспективы развития Корпуса будут обсуждаться в двух разделах. В первом разделе мы коснемся в основном проблем, связанных с пополнением Корпуса новыми текстами (и новыми типами текстов), во втором разделе — проблем совершенствования инструментов поиска и программного обеспечения Корпуса.

Стратегии пополнения Корпуса

Основная характеристика всякого корпуса, помимо, конечно, используемых в нем типов разметки, касается количества и качества представленных в корпусе текстов. В настоящее время по мно-

гим параметрам достигнута относительная сбалансированность Национального корпуса русского языка, но имеется ряд лакун, которые мы считаем необходимым устранить в первую очередь. В их число входит:

1) Расширение подкорпуса текстов XIX века.
2) Создание подкорпуса текстов первой половины XX века
3) Создание нескольких новых типов корпусов, в частности, корпуса поэтических текстов, устных текстов и параллельных текстов

Охарактеризуем эти задачи несколько подробнее.

Расширение подкорпуса текстов XIX века. Поскольку проблемам, связанным с созданием корпуса текстов XIX века, в настоящем сборнике посвящена специальная статья Н. Л. Дич, укажем лишь, что до сих пор по причинам скорее технического порядка в Национальном корпусе русского языка из текстов XIX века были доступны практически только произведения художественной литературы (причем русская классическая литература была представлена относительно равномерно и полно). Максимальный объем художественных текстов этого периода можно оценить приблизительно в 30 млн. словоупотреблений. Остальная часть подкорпуса текстов XIX века должна комплектоваться из нехудожественных текстов, и именно эта задача является в ближайшее время актуальной. В частности, представляют особый интерес научные и научно-технические тексты XIX века (по различным областям знания), публицистика, юридические документы, а также частная переписка, дневники и другие тексты. В целом для подкорпуса XIX века актуальная задача максимального жанрового сближения с подкорпусом современных текстов — разумеется, в пределах возможного. В настоящее время эти два подкорпуса более или менее сопоставимы только в отношении художественных текстов, что, конечно, делает подкорпус XIX века существенно менее сбалансированным и менее пригодным для решения целого ряда лингвистических задач (в первую очередь, лексикографических).

Создание подкорпуса текстов первой половины XX века. Эту задачу мы считаем одной из важнейших в ближайшем будущем. Начав работу над созданием Национального корпуса русского языка (по понятным причинам) с современных текстов, мы на всем протяжении этой работы постоянно помнили о необходимости естественного хронологического расширения нашего массива современ-

ных текстов еще на пять десятилетий «вглубь», т. е. на период с 1901 до 1950 гг. Более того, имея в виду всю условность точной хронологической границы, мы в ряде случаев уже включили в наш подкорпус современных текстов тексты более раннего периода. Частично это было сделано в порядке эксперимента, чтобы отладить систему поиска примеров по дате создания текста, а для некоторых тестов исключение было сделано ввиду их особой культурной значимости не только для времени, когда они были созданы, но и для второй половины XX века, когда они играли активную роль в общественно-политическом дискурсе (таковы, в частности, многие тексты М. А. Булгакова). Тем не менее очевидно, что тексты первой половины XX века должны быть представлены в Корпусе в максимальной полноте, поскольку это период является важнейшим для русского языка и русской культуры. Именно в этот период были созданы многие классические произведения художественной литературы; кроме того, именно этот период характеризуется несколькими уникальными особенностями. С одной стороны, это период резких языковых изменений, вызванных революцией 1917 г. и последующей сменой социального строя. Те языковые новации, которые возникли в этот период, очень важны для истории русского языка (независимо даже от их дальнейшей судьбы, так как одним из них суждено было сыграть роль маргинальных экспериментов, тогда как другие закрепились в литературном языке уже на правах нормативных).

С другой стороны, именно в этот период в русском языке возникает уникальная ситуация разделения на «основной» и «зарубежный» массив текстов. Хорошо известно, что эмиграция «первой волны» оставила обширное и значимое культурное наследие, сопоставимое с наследием метрополии (последующие периоды существования русского языка вне России тоже интересны, но всё же отличаются меньшими масштабами[1]). В Корпус необходимо как можно шире включить тексты, связанные с русской послереволюционной эмиграцией — и произведения художественной литературы, и публицистику, и мемуары, и в особенности (как и в отношении других периодов) тексты, не предназначенные для публика-

[1] Заметим, что в подкорпусе современных текстов Национального корпуса русского языка тексты, созданные вне России, присутствуют — это и зарубежная русскоязычная пресса, и художественные произведения писателей второй половины XX века, работавших в эмиграции (от В. П. Аксенова до А. И. Солженицына).

ции, но отражающие многие особенности языка эпохи гораздо полнее, — частные письма и дневники.

Создание новых типов корпусов. Из числа новых проектов такого рода в первую очередь заслуживает упоминания корпус поэтических текстов. Составители Национального корпуса русского языка прекрасно понимали важность присутствия в этом электронном собрании поэтических текстов наряду с прозаическими (и драматическими). До сих пор это не было сделано отнюдь не из-за недооценки важности поэтических текстов для изучения языка[2], а из-за особых технических сложностей их обработки и разметки. Поэтические тексты было бы желательно включать в корпус не только с морфологической и метатекстовой разметкой — целесообразно было бы сразу снабдить их и особой разметкой, учитывающей важнейшие параметры русской метрики, строфики, рифмовки и т. п. Решить эту задачу одновременно с созданием основного массива Национального корпуса не было возможности, поэтому задача создания корпуса русских поэтических текстов становится особенно актуальной на следующем этапе.

Из других корпусов рассматривается возможность создания особого подкорпуса устных текстов (основой для него будут аудиозаписи, сделанные непосредственно для Национального корпуса в Москве и других городах России) и так называемого «мультимедийного подкорпуса»[3]. Кроме того, важным является проект создания подкорпуса ***диалектных текстов***, представляющий современную диалектную речь по возможности всех основных географических зон России, по которым имеются соответствующие данные. На первом этапе планируется обработать образцы устных текстов, представляющие от 8 до 12 различных русских диалектов общим объемом около 100 тыс. словоупотреблений[4].

---

[2] Напомним, что именно в поэтических текстах, как известно, наблюдается особенно большая концентрация двух типов явлений: с одной стороны, это «законсервированные» архаизмы, с другой стороны — инновации, не проникающие в нормативную письменную речь вовсе или проникающие в очень малых дозах. Русская силлабо-тоническая поэзия дает также бесценный материал для изучения акцентных (и других просодических) явлений.

[3] Особенностям работы над этими корпусами посвящены статьи Е. А. Гришиной в настоящем сборнике.

[4] Об особенностях работы над этим небольшим (но важным) корпусом см. статью А. Б. Летучего в настоящем сборнике.

Разумеется, будет продолжаться работа и над другими корпусами, уже размещенными на сайте Национального корпуса русского языка: это и параллельный англо-русский корпус (см. статью Д. О. Добровольского и др. в настоящем сборнике), и экспериментальный корпус с синтаксической разметкой, и корпус ранних древнерусских текстов.

Совершенствование инструментов поиска и программного обеспечения Корпуса

Одним из основных показателей ценности Корпуса является богатство поисковых возможностей, предоставляемых пользователю. В этой области создателям Национального корпуса русского языка хотелось бы усовершенствовать очень и очень многое. Опять-таки, здесь, как и в предыдущих случаях, перечислим не все вообще мыслимые (и немыслимые) пожелания такого рода, а лишь те, которые намечены для непосредственного воплощения в жизнь в ближайшее время — лишь изредка позволяя себе порассуждать о чём-то большем.

В первую очередь, необходимо усовершенствовать комплекс средств, позволяющих получать разного рода ***статистические сведения*** о Корпусе. В настоящее время в Корпусе существует так называемая «статическая статистика» (указывающая общее количество словоформ на некоторый фиксированный момент времени, общее количество текстов, распределение текстов по жанрам и т. п.) и так называемая «динамическая статистика» (указывающая, какое количество примеров на интересующее пользователя явление найдено по данному запросу). Оба вида статистики нуждаются в постоянной поддержке и совершенствовании.

Однако работы в области корпусной статистики важны и в другом отношении. Насущной задачей усовершенствования разметки Корпуса является разработка и «обучение» программ автоматического снятия грамматической омонимии на основе статистических данных (наподобие тех, что реализованы, например, в Чешском национальном корпусе). В настоящее время в Корпусе, как известно, имеется фрагмент, где грамматическая омонимия не снята, и фрагмент (около 4 млн. словоупотреблений, к началу 2005 года объем этого фрагмента достигнет 5 млн.), где грамматическая омонимия снималась вручную. Разработка программ автоматического снятия грамматической омонимии позволит добавить к этим

двум фрагментам подкорпус, в котором грамматическая омонимия будет снята автоматически с небольшим количеством ошибок (стандартная погрешность в существующих образцах колеблется в пределах 3-5%).

Кроме того, совершенно очевидно, что пользователь должен иметь возможность получить сведения о статистическом распределении тех или иных грамматических, лексических, семантических и прочих явлений по годам, т. е. возникает необходимость в создании интерфейса выдачи *«хронологической» статистики* по данным Корпуса (все данные для которой — датировка текстов, количество тех или иных лингвистических единиц и категорий в данном конкретном тексте, длина данного текста в словах — уже содержатся в Корпусе, вопрос только в том, как построить удобный для пользователя интерфейс выдачи этой информации, а также в том, чтобы создать программное обеспечение, необходимое для требуемых расчетов). Это позволит, в частности, легко датировать первое вхождение того или иного слова в Корпус, отслеживаться хронологическое распределение отдельных лингвистических явлений, и проч.

Другим важным направлением дальнейшей работы является совершенствование ***акцентного*** компонента подкорпуса со снятой грамматической омонимией. В настоящее время расстановка ударений (и буквы *ё*) производилась в подкорпусе со снятой грамматической омонимией без учета некоторых специальных случаев и индивидуальных отклонений (перенос ударения при энклиноменах на предлог, сдвиг ударения в счетной форме и в сочетаниях типа *ни чертá* [*не смыслить*] и т. п.).

К дополнительным поправкам, своего рода мелкой «настройке» морфологического компонента можно отнести и работу по пополнению словаря Корпуса. Включение в словарь лексем, отсутствующих в рабочих словарях Корпуса (на основе которых действуют программы автоматического морфологического анализа), позволит избежать большого числа паразитических разборов и ошибок в подкорпусе с неснятой омонимией. В наибольшей степени нуждаются в пополнении словарь собственных имен, словарь заимствований (особенно недавних) и словарь сленговых слов, не фиксируемых современными нормативными лексикографическими изданиями: это три самых частых источника ошибок морфологического анализатора.

Необходимо также совершенствование «сервисного» программного обеспечения Корпуса, в особенности ***параметров выдачи*** примеров пользователю. В настоящее время в Корпусе действует только так называемый «псевдослучайный» принцип выдачи примеров, согласно которому примеры выдаются в произвольном порядке. При небольшом числе примеров эта проблема не слишком важна, но с увеличением объема Корпуса отсутствие возможности «настраивать» меню выдачи примеров создает определенные неудобства. Во многих случаях пользователь заинтересован в том, чтобы начинать просмотр с определенной части примеров, обладающих теми или иными свойствами — например, созданных до или после определенной даты, принадлежащих определенному жанру или определенному автору, и т. п. Все эти возможности планируется интегрировать в меню поиска и выдачи примеров. Кроме того, планируется ввести и так называемую выдачу по типу KWIC, т. е. упорядоченную относительно правого и/или левого контекста запроса: это стандартный способ выдачи примеров в мировой практике организации корпусов, принятый, в частности, в Британском и Чешском корпусах. В Национальном корпусе русского языка выдача KWIC реализована только в параллельном корпусе англо-русских текстов; в основном же корпусе она пока отсутствует. Наличие детальной морфологической разметки в Национальном корпусе русского языка отчасти компенсирует отсутствие модели KWIC, но возможности, которые предоставляет этот способ сортировки, полезны для решения многих лингвистических задач. Не следует забывать и о том, что KWIC обеспечит лучшую «совместимости» НКРЯ с другими крупными корпусами, используемыми в настоящее время.

Из более частных проблем выдачи информации следует упомянуть возможность сортировки данных Корпуса по социологическим параметрам[5]. Разработка необходимого программного обеспечения позволит в большей мере, чем сейчас, обеспечивать исследовательские интересы социолингвистов, в частности, на более прочном фундаменте проводить гендерные исследования по данным Корпуса (сейчас социолингвистические изыскания возможны только по данным метаразметки основного Корпуса,

[5] О социологической разметке в устном подкорпусе см. с. 108 настоящего сборника.

что, к сожалению, предоставляет социолингвисту лишь достаточно опосредованные данные).

В последнее время разработчики Корпуса обсуждают необходимость и возможность создания «последовательно сужающейся» выдачи примеров из Корпуса по тому или иному запросу. Предположим, например, что пользователю требуется исследовать употребление в русском языке глаголов с приставкой *раз-* и постфиксом *-ся*. Сейчас он может решить эту проблему достаточно неудобным способом: 1) запросить все контексты, в которых используются глаголы с приставкой *раз-* (запрос: V + раз*) или с постфиксом *-ся* (запрос V + *ся) и 2) вручную отбирать те случаи, которые его интересуют. Понятно, что при громадной частоте обеих морфем объем ручной работы в этом случае превосходит все разумные пределы. Очевидно, что было бы гораздо удобнее для выполнения этой задачи 1) запросить все контексты употребления глаголов с одной из двух морфем, а потом 2) на полученном массиве, с использованием функции «искать в найденном», сделать запрос на глаголы со второй морфемой. Аналогичных ситуаций при работе с Корпусом возникает достаточно много, что, очевидно, требует от разработчиков Корпуса совершенствования программного обеспечния в этой зоне.

Дальнейшего усовершенствования требует и метаразметка. В частности, планируется ввести дополнительный набор помет, касающихся «стилистической» характеристики текста, т. е. задающих деление текстов на стилистические нейтральные (основной массив), сниженные (отличающиеся использованием повышенно экспрессивной и грубой лексики) и экспериментальные (отличающиеся обилием «авторских» неологизмов). Такое деление важно прежде всего для настройки лексикографического поиска: в текстах последних двух категорий резко увеличена доля «несловарных» лексем, что может оказаться для пользователя как нежелательным препятствием, так и, напротив, важным преимуществом — в зависимости от его целей. Наличие «стилистического фильтра» позволит пользователю легко исключать стилистически маркированные тексты (или, напротив, работать только с ними).

Наконец, существенный прогресс может быть достигнут уже в ближайшее время в области морфо-семантической и словообразовательной разметки, элементы которой в настоящее время реализованы в Корпусе в экспериментальном порядке. Здесь прежде

всего необходимо добиться, чтобы семантический поиск мог осуществляться с учетом разделения лексемы на значения, в частности, только по «первому» (или основному) значению лексемы. В таком случае в запросе по признаку «музыкальный инструмент» будут выданы примеры со словами *контрабас* и *гитара*, но не примеры со словом *тарелка* или *труба*, способные создать избыточный «шум». Для оптимизации программ семантического поиска необходимо частичное снятие лексической неоднозначности в Корпусе; подходы к решению этой сложнейшей задачи (например, с помощью системы автоматических фильтров, основанных на сочетаемостных свойствах лексем) в настоящее время активно разрабатываются[6].

Разумеется, существует еще много других задач, требующих решения[7]. Более того, пополнение Корпуса часто вызывает к жизни такие проблемы, о существовании которых составители не догадывались, приступая к этой работе. (А если бы догадывались, то кто знает, был ли бы Корпус тогда вообще сделан?)

[6] Более подробно об этом можно прочесть в статье Г. И. Кустовой и др. в настоящем сборнике.

[7] Среди прочего следует упомянуть задачу удобного доступа пользователя Корпуса к разнообразной лексикографической информации, а именно — возможность немедленного получения данных об уже существующей фиксации того или иного слова в разнообразных словарях русского языка.