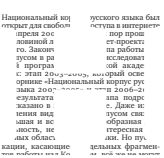
Е. В. Рахилина

# *Корпус как творческий проект*

## Введение



Национальный корпус русского языка был открыт для свобод-
ного доступа в интернете 29 апреля 2004 года — с тех пор прошло 5 с половиной лет, для интернет-проекта это много. Закончились два этапа работы над корпусом в рамках особой исследовательской программы Российской академии наук: этап 2003–2005, который освещен в сборнике «Национальный корпус русского языка 2003–2005» и этап 2006–2008. О результатах второго этапа подробно рассказано в этом сборнике. Даже из оглавления видно, что с Корпусом связана большая и всё более разнообразная деятельность, несомненно, интересная для разных областей лингвистики. Но публикации, касающиеся отдельных фрагментов работы над Корпусом, всё же не могут дать представления о проекте в целом, его развитии, общих задачах и перспективах, его, если можно так сказать, «философии». Восполнить этот пробел мы и попробуем в настоящей статье.

Прежде всего, напомним, что первый этап работы был нацелен на создание корпуса как такового: нужно было собрать как можно больше текстов, сделать корпус представительным и организовать по имеющимся текстам хотя бы самый простой поиск. Все усилия разработчиков были направлены именно на это. Имелось в виду, что главной задачей является «канонический» сбалансированный стомиллионный корпус современного русского языка, хронологические границы которого задавались периодом с 50-х годов XX века по настоящее время. Дополнительно предполагался корпус XIX и первой половины XX века в качестве, так сказать, диахронической составляющей. Все другие разработки, касающиеся диалектного корпуса, корпуса устных текстов, параллельного корпуса и проч. на первом этапе представлялись как экспериментальные, они создавали задел на будущее. Сами эти корпуса в то время либо отсутствовали, либо были очень малы, но активно обсуждались принципы их формирования, их структура, поисковые возможности и т.п. Кроме того, в рамках НКРЯ развивались еще два самостоятельных больших корпусных проекта: корпус XI–XIV вв. и синтаксически размеченный корпус современного русского языка. Работа над первым частично отражена в статье А. И. Зобнина и А. В. Сахаровой в настоящем сборнике; о втором проекте можно прочитать в [Апресян и др. 2005], а воспользоваться этим подкорпусом и изучить принятую в нем систему разметки можно теперь непосредственно на сайте НКРЯ (*http://ruscorpora.ru/search-syntax.html*).

Задачи первого этапа удалось выполнить почти все; собственно, тогда сил не хватило только на систематический сбор текстов первой половины XX века, поэтому данная часть работы завершается только сейчас. В остальном, к 2005 году Национальный корпус русского языка действительно существовал в довольно солидном объеме: 100 млн словоупотреблений, как и планировалось, для современного русского языка и более 20 млн словоупотреблений — для (в основном художественных) текстов XIX века. На этих текстовых массивах уже тогда работал морфологический анализ и пилотный проект семантической разметки. Кроме того, был создан значительный по объему (более 4 млн словоупотреблений) корпус со снятой вручную грамматической омонимией, который давал возможность высокоточной выдачи результатов по запросам,

учитывающим грамматические характеристики лексем. Казалось бы — что еще нужно?

Но нужно еще очень многое. Ведь совокупность существующих на русском языке текстов очень значительна как в пространстве, так и во времени. В Национальном корпусе нужно отражать и все хронологические срезы языка, и все его региональные, социальные и прочие варианты, а вариативность по этим параметрам в русском языке, как известно, достаточно велика. Полноценное отражение такой вариативности — это *первая* задача.

В некоторых случаях варианты превращаются почти что в отдельные подъязыки, для которых нужно строить свои подкорпуса со своей специально настроенной на них системой разметки. Игнорировать такие слои русского языка никак нельзя: чем сложнее они устроены, тем больше их значимость для системы в целом. Значит, это *вторая* задача.

*Третья* задача неожиданно обнаружилась непосредственно во время работы над Корпусом. Разработчики и разметчики трудились с таким энтузиазмом, что объемы корпуса росли стремительно — и уже к концу первого этапа старые технологии не могли справиться с ними. Корпус стал работать медленно и с перебоями, отказываясь «отвечать» на сложные запросы. Понадобилось его «техническое перевооружение».

*Четвертая* задача — популяризация Корпуса. К 2005 году основными его пользователями оставались иностранные слависты, которые, во-первых, уже привыкли к работе с корпусами других европейских языков, а во-вторых, получили огромный открытый ресурс, позволяющий относительно объективно оценивать правильность или распространенность тех или иных форм или конструкций русского языка, не прибегая к трудоемкой «человеческой» экспертизе. Между тем, конечно, Корпус нужен в России и делался прежде всего для русскоязычных пользователей — и лингвистов, и не только лингвистов. Например, для нового поколения учащихся компьютерные продукты уже не менее привычны, чем книги, и если мы хотим сохранить интерес к русскому языку в следующих поколениях, нужно думать об этом сегодня. Но для того, чтобы Корпус стал доступен широкому кругу пользователей — от школьников и школьных учителей до любителей русского языка в любой точке

нашей страны — нужна большая работа. Это, с одной стороны, работа просветительская, а с другой — техническая: оснащение Корпуса разнообразными пользовательскими инструкциями, подкорпусами с упрощенной (или, наоборот, со сложной специальной) разметкой, введение поисковых настроек, которые бы облегчали его использование, и т.п.

И, наконец, *пятая* задача — широкое использование корпуса для построения на его базе новых лингвистических продуктов: новых словарей и новых грамматических описаний. То есть, собственно, то, для чего всякий корпус и создается.

Вот эти пять задач и описывают программу развития Национального корпуса русского языка. Теперь по порядку о том, как они решались в 2006–2008 годах и что предполагается в этом плане делать дальше.

## 2. Пополнение Корпуса

Итак, речь идет о хронологических (2.1), пространственных (2.2) и социальных (2.3) срезах. Что здесь сделано — и что еще предстоит сделать?

**2.1.** В период 2006–2008 гг. в Институте русского языка им. В. В. Виноградова совместно с Казанским государственным университетом начата работа по созданию подкорпуса XVIII века (см. подробнее статью С. О. Савчук и Д. В. Сичинавы в настоящем сборнике). Таким образом, с учетом корпуса XIX века (см. статью С. А. Оскольской), корпуса первой половины XX века (см. статью С. О. Савчук) и основного корпуса в перспективе речь идет об охвате фактически всего периода существования современного русского литературного языка. Создание и обработка обширной (более двух миллионов словоупотреблений) коллекции текстов XVIII века — это важный шаг, потребовавший значительных усилий, потому что в этой временной зоне разметчики сталкиваются с существенно более высокой вариативностью по сравнению со стандартным литературным языком, и доля их ручного труда сопоставима с обработкой диалектных текстов. Но и результаты этой работы заметны: благодаря ей уже сейчас в НКРЯ есть возможность мониторинга изменений лексической семантики и синтаксиса. Например, если прослеживать примеры хронологически, видно, что прилагательное *противный* именно на этом отрезке времени начало менять

свою семантику с ‘противоположный’ (*противный берег*) на ‘имеющий отрицательную оценку’ (*противный мальчишка*).

Теперь о том, что еще хотелось бы сделать.

Во-первых, следовало бы пополнить Корпус текстами первой половины XVIII века. Пока в Корпусе присутствуют за небольшими исключениями только тексты второй половины — их несколько легче обрабатывать и они доступнее в электронном виде, поэтому начали с них. Добавление более ранних текстов придаст законченность нашей коллекции литературных текстов и, как мы надеемся, вдохновит историков языка на «встречное» движение — создание близких по времени корпусов позднего среднерусского периода XVI–XVII вв.

Во-вторых, конечно, в ближайшие годы нужно пополнить и основной корпус, который «остановился» на 2005 г., так что нужна сбалансированная подборка и более поздних текстов, скажем, до 2010 г. Но объем основного корпуса при этом, видимо, должен остаться старым — 100 млн словоупотреблений. Один из возможных вариантов решения этой проблемы — удалить из Корпуса какое-то количество набранных ранее текстов и с этой целью образовать Банк Корпуса, в котором хранились бы (и были доступны) все «лишние» тексты.

В-третьих, нужно продолжать работу по созданию качественно сбалансированных коллекций по всем периодам. Действительно, когда работа только начиналась, баланс соблюдался условно. Например, понятно, что и для периода XIX века, и для большей части XX-го художественная литература более доступна, чем публицистика, а тем более частная переписка и другие маргинальные жанры. Естественно, что акцент делался на как можно более полный охват художественной литературы. Но в условиях, когда срез устной речи отсутствует полностью, и публицистика, и эпистолярный или дневниковый жанр оказываются крайне важны для корпуса, потому что они отражают несколько другой — по сравнению с литературно-художественными текстами — вариант языка, более близкий к повседневному разговорному языку того времени. Значит, нужно и дальше искать, обрабатывать и вводить в Корпус новые тексты этих жанров для соответствующих временных периодов.

**2.2.** Если не считать диалектного подкорпуса, то пространственные срезы русского языка пока представлены в НКРЯ только в периодике основного корпуса, где есть региональные газеты, и в устном корпусе — благодаря хрестоматиям (таким, как [Сергеева, Герд (ред.) 1998]). В перспективе, конечно, тут нужна большая работа прежде всего по сбору материала в разных регионах России, на постсоветском пространстве, а также речи эмигрантов разных поколений[1]: фрагменты таких текстов обязательно должны быть включены в Корпус.

Что касается регионов России, то эта задача крайне насущная, и требует она не столько больших денег или усилий, сколько доброй воли лингвистов в регионах: ведь не секрет, что в самых разных университетах (в Перми, Омске, Барнауле, Томске, Челябинске и мн. др.) ведется сбор и коллекционирование устных текстов в рамках различных программ и проектов и просто студенческой практики. В отсутствие единого координационного центра эти тексты в лучшем случае выходят в виде хрестоматий, но оказываются недоступны электронно, обычно же — вкладываются в виде отдельных примеров в малотиражные монографии или диссертации, которые трудно получить уже не только в электронном, но и в бумажном виде, чаще всего же они просто теряются и пропадают. Добрая воля соответствующих кафедр, лабораторий и самих исследователей регионального разговорного языка и просторечия состояла бы в том, чтобы — параллельно с использованием в диссертациях, монографиях и хрестоматиях — эти материалы предоставлялись в Корпус, где бы они обрабатывались и становились общедоступными при поиске, в соответствии с законом об авторском праве, отдельными фрагментами — конечно, со всеми необходимыми ссылками, благодарностями и письменными обязательствами о нераспространении целых текстов, как это принято в НКРЯ. Пока так сотрудничают с Корпусом Саратовский университет — известная группа О. Б. Сиротининой, русская кафедра Хельсинкского университета (ее представляет Е. Ю. Протасова) и — пока, так сказать, в пилотном формате — Петербургский уни-

[1] Один из примеров такого собрания (и одновременно его анализа) — книга Е. Ю. Протасовой [2004].

верситет (М. В. Русакова и лаборатория А. С. Асиновского). Мы искренне благодарны этим коллективам и надеемся на то, что этот удачный опыт обретет последователей.

**2.3.** Теперь о работе над представлением в Корпусе социально значимых срезов русского языка. Наибольший объем работы за период 2006–2008 гг. выполнен в области устных текстов — в результате для русского языка фактически создан и функционирует полноценный (5,5 млн) подкорпус устной речи с особой системой разметки (подробнее см. статью Е. А. Гришиной и С. О. Савчук в наст. сб.), в частности, отражающей гендерные различия говорящих, который по объему превосходит, например, японские аналоги (см. статью А. В. Костыркина). Причем, если японские тексты записаны в студийном формате, русские, в значительной своей части, собраны, говоря языком лингвистов, «в поле» — т.е. представляют собой живую спонтанную речь и, тем самым, с лингвистической точки зрения, обладают повышенной ценностью (в Корпус включены как прежние, ранее собранные различными исследователями и уже опубликованные записи устной речи жителей Москвы, С.-Петербурга и других городов, так и записи, полученные непосредственно составителями Корпуса). Другой особенностью этого подкорпуса является коллекция кинофильмов, вручную и с большой степенью подробности расшифрованных группой Е. А. Гришиной. Аналоги корпусному кино-проекту нам неизвестны. Но, конечно — и об этом мы только что говорили в предыдущем разделе — устный подкорпус, для качественных и количественных характеристик которого задана такая высокая планка, не должен стоять на месте, и мы надеемся на его продолжение и развитие (см. 3.1).

Другой важный проект — это тексты электронной коммуникации. Здесь работа только начинается и требует больших затрат, потому что интернет-тексты создаются с нарушением орфографической и грамматической правильности, содержат большую вариативность и фактически нуждаются в особом словаре. Но лингвистически это очень важный пласт языка, потому что именно здесь происходят инновационные процессы, причем несколько иные, чем в разговорной речи. Во-первых, среди электронных текстов много узкоспециальных, со своей терминологией: форумы автолюбителей, футбольных фанатов и т.п. Во-вторых, это, хоть и особые, но

все-таки письменные тексты, а значит, в них вырабатываются свои правила письма — и в области орфографии, и в области организации дискурса. Будут ли эти правила затем влиять на общелитературную речь? Или, может быть, уже влияют? Все это нуждается в скорейшем изучении, но для квалифицированного ответа на такие вопросы нужен современный и достоверный источник данных, которым должен быть постоянно пополняемый корпус с итоговым объемом не менее 5 млн словоупотреблений.

### 3. Специальные подкорпуса: устный и медиа-, диалектный, поэтический, акцентологический, параллельный

**3.1. Устный подкорпус.** Подкорпус в Корпусе выделяется тогда, когда ему соответствует не просто особая коллекция текстов, связанных общими свойствами (например, временными рамками), но и особая система помет. Теперь так устроен корпус устных текстов: в процессе развития в период 2006–2008 гг. он выделился в отдельный ресурс, хотя еще и остается «похож» на основной корпус. Если всё будет развиваться так, как мы сегодня планируем, в ближайшем будущем его ждут большие перемены, которые вначале коснутся только его фрагмента — киноколлекции. Она перерастет в Мультимедийный русский корпус, или МуРКо, и обретет звуковой и видеоряд (подробнее об этом проекте см. статью Е. А. Гришиной «Мультимедийный русский корпус (мурко): проблемы аннотации»). С точки зрения всей программы развития Корпуса, это был бы важный результат, поскольку для его достижения неизбежно потребуется внедрение и отработка новых технологий. Ведь звуковая и видеодорожки — это не просто механическое расширение объема Корпуса, а прежде всего возможность соотнести речевой или видеофрагмент с его письменной записью, организовать по ним поиск. Если «испытание» пилотного проекта пройдет успешно, затем, так сказать, по следам устного корпуса, те же технические решения можно будет применять и к другим фрагментам нкря — например, в добавлении звуковой дорожки остро нуждается диалектный подкорпус.

**3.2. Диалектный подкорпус.** Диалектный подкорпус представляется как часть Национального корпуса русского языка — но, конечно, особая часть. Он очень маленький — в 100 с лишним раз меньше нкря, но он требует гораздо более сложной разметки (см.

статью А. Б. Летучего в настоящем сборнике) и более трудоемкой ручной обработки, чем обычные тексты. Корпус проектировался и создавался с ориентацией на, так сказать, рядовых пользователей корпуса, большинство из которых никогда в жизни не видело ни одного диалектного текста. В то время задачей было сделать своего рода «научную игрушку», которая наглядно демонстрировала бы разнообразие русского языка в его региональных вариантах.

Корпус создавался при активном содействии диалектологов — прежде всего, Саратовской группы В. Е. Гольдина и специалистов из Московского государственного университета. Однако в массе своей диалектологи к этой идее относились с опаской (впрочем, как поначалу и держатели всех других типов уникальных текстов — к Корпусу вообще): не окажется ли эта идея пустым и бесполезным делом? Однако уже первая работающая версия диалектного корпуса породила огромный энтузиазм, и подкорпус стал получать «добровольные пожертвования» в виде электронных текстов, записанных исследователями самых разных диалектологических центров России — Курска, С.-Петербурга, Волгограда и многих других. Одновременно пришло понимание, что этот проект полезно было бы перестроить так, чтобы он служил самим диалектологам — и как удобно организованный ресурс для учебного процесса, и как инструмент для исследовательской деятельности. Правда, тогда все технологические и организационные решения должны находиться под контролем заказчиков, потому что специалисту-диалектологу от Корпуса нужно гораздо больше, чем обычному пользователю. В частности, диалектологи хотели бы видеть здесь свои собственные фонетические записи текстов, а не только тот упрощенный вариант унифицированной орфографической транскрипции, который сейчас делает возможным поиск одновременно по всему массиву разнообразных диалектных текстов, — очевидно, что в Корпусе нужна и возможность поиска, и подлинная фонетическая запись. По-видимому, для новых задач понадобится и уточнение транскрипции, есть мечта добавить звуковую дорожку — словом, обнаружилось, что этот проект чрезвычайно востребован и его необходимо развивать.

Сейчас «перестройка» диалектного корпуса находится в стадии творческого обсуждения — мечтаний, споров, проб и даже, навер-

ное, ошибок; постепенно эта работа войдет в общее русло — и мы все очень надеемся на ее успех.

**3.3. Поэтический корпус.** До 2005 года Национальный корпус русского языка говорил прозой, а между тем, русская литература и русский язык немыслимы без русской поэзии. Конечно, можно было бы «забыть», что стихи — это стихи, но разработчики пошли другим путем и за три года создали новый продукт: поэтический подкорпус с уникальной системой разметки и поиска (подробнее см. статью Е. А. Гришиной, К. М. Корчагина, В. А. Плунгяна и Д. В. Сичинавы в наст. сб.), аналогов которой, насколько нам известно, нет в мире (как нет и других поэтических корпусов).

В настоящее время этот подкорпус охватывает XVIII и XIX век, а также некоторых поэтов начала XX века. Если говорить о развитии — то для этого подкорпуса задача формулируется очень просто: увеличение объема, и мы надеемся охватить хотя бы классическую поэзию XX века (условно — до Бродского и Окуджавы), а в идеале включить всё, включая тексты популярных песен и рок-поэзию. (Правда, чем дальше, тем сложнее работа: уже поэты конца XIX в. требуют более сложной обработки, чем авторы классических ямбов или хореев, — что уж говорить об авторах XX века!)

Задач у такого корпуса очень много. Конечно, прежде всего он ориентирован на филологов, которые получают новый инструмент исследования поэтического языка и просто полную электронную коллекцию поэтических текстов (далеко не все из которых были легко доступны). Для преподавателей (даже школьных) — это возможность мгновенно получить большое число примеров на разные типы стихотворного размера, а для исследователей-стиховедов — компактный и эффективный справочник по русской метрике, рифме, строфике и другим параметрам стиха. В целом же в рамках этого проекта речь идет не просто о сохранении русского языка или литературы, но о поддержании целого пласта, в общем, исчезающих культурных традиций.

**3.4. Акцентологический подкорпус.** Русское ударение подвижно, но, как известно, в письменных текстах не ставится — поэтому по ним невозможно восстановить, как действительно был произнесен тот или иной текст. Конечно, есть правила, регламентирующие расстановку ударений — и на основании этих правил в самом

начале работы над Корпусом была построена программа, которая ставит ударение автоматически, правда, только для подкорпуса со снятой омонимией. Но ведь, как известно, реальные говорящие правил не соблюдают — живой язык им диктует свои законы, в том числе и касающиеся ударений, и лингвистам хорошо известно, что схема ударения в слове может меняться. Именно поэтому так важно знать, какие именно отклонения от канонических правил реализуются в сегодняшнем языке и существовали в его предшествующие периоды.

Для современного русского языка установить это можно, акцентуируя вручную устные тексты. Для языка прошлых веков — анализируя поэтические строки, в которых метр основан на чередовании ударных и безударных слогов в строке. По мере развития НКРЯ, а с ним и двух новых подкорпусов — устного и поэтического — все более реальной становилась идея создания специального исторического акцентологического подкорпуса, объединяющего поэтический и устный подкорпус (прежде всего, кинотранскриптов) с проставленным вручную ударением. Идея (как всегда, совершенно нестандартная) принадлежит Е. А. Гришиной, она является организатором и главным исполнителем всего этого проекта (см. ее статью «Корпус “История русского ударения”» в наст. сб.). Сам проект только начался, но его первые результаты можно уже сейчас увидеть на сайте Корпуса.

**3.5. Параллельный подкорпус.** В том виде, в котором он сейчас представлен в НКРЯ, параллельный корпус начинался как совместный проект ИРЯ им. В. В. Виноградова РАН и Воронежского государственного университета. К 2005 году в порядке эксперимента был обработан корпус переводов с русского на английский и с английского на русский объемом свыше полутора миллионов словоупотреблений (подробнее об этой работе см. [Добровольский и др. 2005]). Эти тексты имели совершенно другой формат представления, чем тот, который был свойствен НКРЯ в целом, поэтому они не могли быть размещены на том же сайте и не могли получить ту же разметку, что и остальные тексты Корпуса. В результате, при поддержке С. А. Шарова, которому мы очень благодарны за содействие, наш параллельный корпус был размещен на сайте университета г. Лидс (Великобритания). Однако, как показала практика,

такое «дистантное» управление корпусом не очень удобно, и нашей мечтой было уговорить программистов компании «Яндекс» адаптировать этот ресурс к возможностям нашего сайта. Корпус рос и развивался, но мечта всё не сбывалась. Разработчики уже начали новый эксперимент: немецко-русский параллельный корпус, но и его приходилось отправлять в Англию.

И вот, в этом году, в связи с общей технической перестройкой НКРЯ, задача перевода параллельного корпуса на «Яндекс» наконец-то была решена. При этом потребовался перерыв в его работе на полгода — зато теперь в английской составляющей корпуса работает не только лексический, но и морфологический поиск, и при этом для запросов доступен весь тот материал, который был накоплен за прошедшие годы — более 7,5 млн. в англо-русской и свыше 1,5 млн. — в русско-английской части корпуса.

Теперь, когда параллельный корпус сопряжен с основным, хочется думать о его серьезном дальнейшем развитии. Востребованность параллельных корпусов очень высока. Причем если англо-русский и русско-английский корпуса, равно как и аналогичный немецко-русский ресурс, нужны прежде всего для оптимизации методик обучения иностранному языку, то — шире — выровненные тексты вообще могут и должны служить базой для различных типологических исследований. Поэтому, как отмечалось, например, на последнем — XIV — съезде славистов, высока потребность в параллельных русско-славянских корпусах, в частности, ориентированных на польский, чешский, болгарский, словенский и др. языки. Необходимость в подобных ресурсах есть даже для очень близких пар — таких, как русский и украинский или русский и белорусский. Другое направление развития параллельных корпусов связано с созданием многоязычных ресурсов. Зачастую они включают в себя выровненные переводы одного художественного произведения на различные языки. Над корпусами такого рода сегодня активно работают известные типологи многих стран (ср. проекты И. ван дер Ауверы в Бельгии, Т. Штольца в Германии, А. Барентсена в Нидерландах и др.). Классическими объектами этой работы являются «Маленький принц», «Гарри Поттер» и «Алиса в стране чудес». В большинстве случаев результаты таких проектов не могут пока свободно распространяться, так как неограниченный ин-

тернет-доступ к полному тексту произведения в настоящее время нарушает авторские права. Однако отработанные уже технологии НКРЯ позволяют выдавать текст небольшими фрагментами, а значит, у нас есть принципиальная возможность сделать такой корпус общедоступным. Осталось ее реализовать.

### 4. «Техническое перевооружение» Корпуса

История этого вопроса такова: в 2005 году, на следующий год после того, как была сдана и вывешена в интернете первая очередь Корпуса, который к тому времени как раз перевалил за стомиллионный объем и был размечен не только морфологически, но и семантически, мы впервые столкнулись с серьезными перебоями в его работе — происходило то, что на жаргоне программистов называется «корпус упал». Это проявлялось в том, что на сколько-нибудь сложные запросы (неоднословные, с участием морфологической, а тем более семантической информации) пользователь получал быстрый и лаконичный ответ о невозможности выдать результаты из-за нехватки памяти. Нужно было срочно менять формат представления данных (переходить с HTML на XML), увеличивать объем и быстродействие сервера и вообще совершенствовать корпусные технологии — этап «технического перевооружения» был произведен благодаря специалистам компании «Яндекс» (некоторые детали этого процесса изложены в статье А. А. Аброскина в настоящем сборнике), причем на это потребовалось довольно много усилий и времени: несмотря на то, что уже давно нет сбоев в функционировании сервера, работа над решением некоторых насущных задач все еще продолжается.

Между тем, благодаря такой «технической перестройке» в Корпусе появилось много новых полезных функций — например, при поиске стало возможным учитывать знаки препинания (в том числе искать слово до или после запятой, точки или, скажем, вопросительного знака), а также учитывать регистр — заглавные или строчные буквы.

Кроме того, наконец, разрешилась известная проблема KWIC-выдачи. Дело в том, что в широко принятом в корпусной лингвистике формате — так сказать, в корпусном стандарте — положено, чтобы у пользователя была возможность на запрос о слове видеть

его правый и левый контексты. Обычно для этого используется такой вид страницы, при котором все выданные в ответ на запрос предложения центрируются, причем центральным (и зрительно выделенным) оказывается запрошенное слово, а его правый и левый контекст в каждом предложении отделены от него дополнительными пробелами. Таким образом, страница выдачи выглядит как столбик одинаковых слов, каждому из которых слева на некотором расстоянии приписаны непосредственно предшествующие ему фрагменты контекста, а справа — следующие за ним слова. Предложения видны пользователю не целиком — удобство в том, чтобы сразу просматривать ближайший контекст, по которому к тому же можно производить сортировку. На сайте «Яндекса» нет возможности добиться выдачи в таком формате. Зато мы получили возможность упорядочивать примеры, причем не только по правому и левому контексту (почти KWIC-выдача!), да еще с учетом формы искомого слова, но и по автору, а главное — по хронологии, а это в разы сокращает труд лингвиста по мониторингу изменения тех или иных языковых характеристик во времени.

Особая проблема — разработка и внедрение программы, позволяющей снимать морфологическую омонимию в Корпусе на основе статистических методов. Эта программа была создана для Корпуса уже несколько лет назад (ее автор — А. В. Сокирко), и она тестировалась на нашем корпусе со снятой омонимией. Однако при ее тестировании выявился ряд существенных недочетов, которые, в частности, свидетельствовали об ошибках в тренировочном корпусе. Эти ошибки возникали и по случайным причинам (естественно, что, работая на массиве в несколько миллионов словоупотреблений, разметчики не могут не ошибаться), так и в результате некоторых системных сбоев (например, при смене программ обработки текстов).

Поэтому в 2008 году было принято решение перенаправить те силы и средства, которые были предназначены для увеличения объема корпуса со снятой морфологической омонимией, на его правку и оптимизацию; в настоящее время программа А. В. Сокирко проходит новое тестирование — причем отдельно создается ее вариант для современных текстов, и отдельно — для текстов XIX и первой половины XX века. По результатам тестирования в ближайшее вре-

мя будет принято решение об открытии корпусов со статистически снятой омонимией для каждого из этих периодов.
Но, конечно, это еще не все: в техническом отношении Корпус пока еще нуждается в дальнейшей доработке. Нужно иметь возможность представлять на сайте статистику по каждому запросу, нужно совершенствовать выдачу (вплоть до выгрузки ее в формат Excel), нужен английский (а может быть, и французский?) интерфейс и так далее, и так далее. И все это — для того, чтобы открыть возможности Корпуса широкому пользователю.

### 5. Корпус — широкому пользователю

У этой задачи есть два аспекта: первый — чисто просветительский, он связан с тем, чтобы как можно полнее и ярче донести информацию об имеющемся ресурсе до максимального числа потребителей. Второй — более сложный в исполнении: улучшить пользовательский интерфейс и пользовательские характеристики Корпуса так, чтобы повысить его ценность как информационного продукта. Осознав эти две задачи, мы вели работу в обоих направлениях.

Действительно, пока основная масса пользователей Корпуса — ученые-исследователи; огромный резерв здесь составляют преподаватели и учащиеся самых разных уровней — от школ до университетов, подготовительных курсов, курсов усовершенствования или второго высшего образования. Значительный (более чем трехлетний) опыт в этом отношении накоплен на Отделении деловой и политической журналистики Высшей школы экономики в Москве, где Корпус фактически служит активным инструментом обучения грамматике, стилистике, культуре речи и всему комплексу дисциплин, связанных с русским языком (подробнее см. статью Н. Р. Добрушиной в наст. сб.). На основе Корпуса создаются упражнения к занятиям, контрольные работы, по Корпусу даются домашние задания и курсовые работы, составляются методические пособия и вспомогательные словари. Не случайно именно отделение журналистики ВШЭ стало базой для проведения семинаров совместно с Институтом усовершенствования учителей в 2005–2006 гг., а затем двух общероссийских школ-семинаров по обучению Корпусу — весной 2007 при поддержке ВШЭ и осенью 2008 годов при поддержке Министерства образования и науки РФ.

К работе первой Школы была приурочена Международная конференция по использованию НКРЯ, в которой приняли участие, с одной стороны, слависты из Италии, Финляндии, Франции, США, Швейцарии и других стран, а с другой — русисты из самых разных городов России: Воронежа, Читы, Ульяновска, Новгорода и др. Интерес к Корпусу постоянно растет — и среди лингвистов-исследователей, и среди преподавателей русского языка. В августе 2008 года была организована обучающая Школа-семинар в Казани, в 2009 планируется такая же школа в Гродно. Конечно, разработчики читают лекции, доклады и организуют мастер-классы по Корпусу. Только за период с 2006 по 2008 гг. такие выступления прошли в университетах Томска, Киева, Гродно, Алма-Аты, Вильнюса, Афин, Тромсе (Норвегия), Сеула, Нанта (Франция) и многих других, все это требует больших дополнительных усилий, но их все равно недостаточно. Нужен единый центр, который бы помогал организации обучения и аккумулировал все методические и исследовательские работы и проекты на базе Корпуса. В современных условиях это мог бы быть Интернет-портал, функционирующий при корпусном сайте; его разработка станет одной из главных задач на ближайшие годы.

Между тем портал нужен совсем не только для распространения информации о Корпусе (хотя это и важная задача) и даже не только для объединения лингвистов и преподавателей и обмена информацией между ними: сегодня портал нужен и самим разработчикам — для того, чтобы иметь обратную связь с пользователями и быстрее реагировать на новые потребности, которым должен отвечать Корпус.

Пока портала нет — но некоторое, так сказать, «технологическое движение» ресурса к пользователю происходит и сейчас. В частности, в 2007 году был открыт Обучающий подкорпус, ориентированный на школьников старших классов и их учителей. В нем на материале произведений школьной программы по литературе осуществлена разметка, учитывающая требования программы по русскому языку (подробнее об этом проекте см. статью С. О. Савчук и Д. В. Сичинавы в наст. сб.). В развитие Обучающего подкорпуса на сайте размещены инструкции по пользованию Корпусом, начата работа по словообразовательной разметке. В непосредственном контакте с пользователями происходит и правка системы семанти-

ческих помет (подробнее см. статью Е. В. Рахилиной, Г. И. Кустовой, О. Н. Ляшевской, Т. И. Резниковой и О. Ю. Шеманаевой), и работа над корпусным списком устойчивых сочетаний, и внедрение в Корпус фильтров, частично снимающих семантическую омонимию (см. статью Г. И. Кустовой). В то же время, эти работы имеют и самостоятельную ценность: некоторые из них представляют собой лингвистические продукты нового поколения.

### 6. Корпус и новые лингвистические продукты

Действительно, главная задача, на которую в свое время ориентировались разработчики Корпуса, — это повышение точности и представительности языкового материала в основных лингвистических продуктах, т.е. в словарных и грамматических описаниях; теперь наступило время, когда можно приступить к решению этой задачи. Важный вопрос — с чего начать? Если выбирать между длительными, трудоемкими и сложными проектами, как, например, многотомный толковый словарь, и относительно «короткими» разработками, не требующими больших исследовательских коллективов, то начать целесообразнее с последних — именно на них лучше отрабатывать технологии и практические решения.

Следуя этой логике, мы приступили сначала к разработке нового частотного словаря русского языка, а также серии сочетаемостных словарей — словаря устойчивых оборотов, словаря сочетаемости неполнозначных глаголов с абстрактными именами (типа *принять решение*), словаря сочетаемости прилагательных и наречий высокой степени (типа *смертельная усталость / смертельно устал*); запущены проекты еще нескольких сочетаемостных словарей. Такая работа опирается на словарные базы данных Корпуса и может быть выполнена в довольно сжатые сроки. Оптимальный способ представления результатов здесь — компьютерные системы, а не традиционные бумажные издания, хотя в некоторых случаях бумажные версии (например, для частотного словаря) тоже планируются к выпуску.

Особая задача — создание грамматических описаний, базирующихся на корпусных данных; лингвисты во всем мире начинают сознавать важность разработки грамматик, которые ориентируются не на искусственно сконструированные примеры, а на сово-

купность текстов, действительно порожденных носителями языка. «Существующим в языке», в соответствии с этой новой идеологией, должно признаваться в первую очередь то, что (надежно) засвидетельствовано в корпусе данного языка, а не то, что вытекает из зависящих от весьма гибкой интуиции самого лингвиста суждений о грамматической правильности (подробнее об этой проблеме см. [Плунгян 2008]).

Важно, что в исследовательской среде Корпус постепенно становится, так сказать, стандартной материальной базой для работ по русистике. В частности, уже издано несколько сборников [Добрушина (ред.) 2007, Мустайоки и др. (ред.) 2008], которые специально посвящены корпусным исследованиям в лексике и грамматике, ср. также монографию [Князев 2007] и др. В настоящем сборнике также публикуется несколько научных статей хорошо известных лингвистов, которые на разном материале (устного, параллельного, общего корпусов) иллюстрируют возможности приложения данных НКРЯ к лингвистическому описанию (см. статьи М. Д. Воейковой, Д. О. Добровольского, Е. В. Падучевой). Понятно, что все это пробные фрагменты и что усилия по созданию единого описания русского языка нужно объединять: сама по себе это слишком большая задача. Но раз в этой области уже происходят эксперименты, раз на этом пути есть успехи, значит, она будет решена.

## 7. Заключение

Сборник, который открывает эта статья, очень разнородный — потому что работа над Корпусом включает самые разные виды деятельности. Нашей задачей здесь было представить проект как единый, показать, что его разные аспекты (и отражающие их разные разделы сборника) в конечном счете подчинены некоторой общей стратегии. Однако ни данная статья, ни даже сборник в целом, видимо, не могут отразить главное, — то, что было вынесено в заглавие настоящей статьи: Корпус — это творческий проект. Невозможно рассказать об энтузиазме совсем небольшой группы лингвистов, которые, по сути дела, отложив работу над статьями и книгами, спорят на семинарах, снимают омонимию, собирают тексты, размечают, считают, придумывают… Приходят люди в Корпус заниматься разметкой, а где-нибудь через год они уже воплощают собственные

идеи и фактически управляют «своим» подкорпусом. Поэтому Корпус — это не только интернет-продукт, но и творческое сообщество людей, которые работают вместе. Их творческий заряд и воплощается в структуре этой системы, так что сама она максимально (из существующих корпусов) приспособлена для творческого поиска пользователя. Полем для такого поиска является русский язык.

### Литература

Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л., Иомдин Л. Л., Санников А. В., Санников В. З., Сизов В. Г., Цинман Л. Л. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // НКРЯ 2003–2005, с. 193–214.

Добровольский Д. О., Кретов А. А., Шаров С. А. 2005. Корпус параллельных текстов: архитектура и возможности исследования // НКРЯ 2003–2005, с. 263–296.

Добрушина Н. Р. (ред.) Национальный корпус русского языка и проблемы гуманитарного образования. — М.: Теис, 2007.

Князев Ю. П. Грамматическая семантика: Русский язык в типологической перспективе. М.: Языки славянских культур, 2007.

Мустайоки А., Копотев М. В., Бирюлин Л. А., Протасова Е. Ю. (ред.) Инструментарий русистики: корпусные подходы. Slavica Helsingiensia 34. Хельсинки, 2008.

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

Протасова Е. Ю. Феннороссы: жизнь и употребление языка. СПб: «Златоуст», 2004.

Сергеева Н. С., Герд А. С. (ред.) 1998. Русская разговорная речь европейского Северо-Востока России. СПб: СПбГУ.