АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81'322.2-161.1

Д. В. Сичинава

К задаче создания корпусов русского языка

Статья посвящена теоретическим и практическим вопросам создания размеченных корпусов русского языка (массивов текстов на русском языке, сопровожденных лингвистической информацией). Среди теоретических вопросов рассматривается типология корпусов по принципу ширины "охвата" текстов (полный, культурно-репрезентативный, эталонный корпус), концепция разметки (лингвистической информации в корпусе). Практическая часть включает сведения о разработке русских корпусов в рамках проекта Центра лингвистической документации, о конкретных методах, о встречающихся трудностях и способах их разрешения.

0. ВВЕДЕНИЕ

Работа с корпусами, т. е. с массивами текстов, представленными в компьютерном виде, давно уже стала одним из основных, если не основным методом лингвистических исследований, с помощью которого могут решаться самые разные задачи. Между тем отечественная лингвистика, как известно, отстаёт в этом отношении от современного состояния зарубежных исследований; созданный еще в 1960-е гг. (и то вне России) Уппсальский корпус русских текстов остаётся, насколько нам известно, единственным завершённым и активно используемым проектом такого рода. Не говоря уже об устарелости его материалов и ограниченности объёма (1 млн словоупотреблений), нужно указать прежде всего на то, что он не является лингвистически размеченным (т. е. в нём не указаны морфологические, синтаксические, семантические свойства тех или иных сегментов текста, что затрудняет поиск по нему), в то время как современная лингвистика оперирует в основном аннотированными (размеченными) корпусами (treebanks). Начиная с 1980–1990-х гг. работа над созданием компьютерных баз данных по русскому языку ведется в рамках Машинного фонда русского языка при Институте русского языка РАН под руководством В. М. Андрющенко (работа отражена в многочисленных публикациях, прежде всего [1]).

С появлением глобальной компьютерной сети Интернет всё больше и больше лингвистических корпусов становятся доступными (полностью или частично) языковедам всего мира: Британский национальный корпус (http://thetis.bl.uk), корпус латинских текстов "Персей" (http://www.perseus.tufts.edu), чешский корпус Карлова университета (http://ucnk.ff.cuni.cz) и др. Для нового русского корпуса желательно, чтобы он был доступен и on-line.

В настоящее время в Институте проблем передачи информации РАН ведётся работа по созданию аннотированного корпуса русских текстов на материале собрания Уппсальского корпуса [2], однако работа по созданию "корпуса ИППИ" далеко не завершена. Некоторые соображения относительно этого проекта мы выскажем в дальнейшем.

В нашей статье излагаются некоторые конкретные соображения по созданию русских корпусов в связи с тем, как их реализация мыслится в рамках проекта "Корпус ЦЛД-МГУ", осуществляемого с 2001 г. общественной организацией ЦЛД (Центр лингвистической документации при Московском центре непрерывного математического образования, руководителем ЦЛД является В. А. Плунгян совместно с Кафедрой теоретической и прикладной лингвистики МГУ им. М. В. Ломоносова и компанией "Яндекс". В проекте участвуют также А. Е. Поляков (НТЦ "Информрегистр"), И. В. Сегалович, В. А. Титов (компания "Яндекс"), С. Ю. Толдова и автор этих строк (филологический факультет МГУ). С первой половины 2002 г. к деятельности по созданию корпусов активно подключился также вновь созданный Отдел лингвистических исследований (ОЛИ) Всероссийского института научной и технической информации (зав. отделом — Е. В. Рахилина).

1. СОДЕРЖАНИЕ КОРПУСОВ

Нынешние исследования в области корпусов имеют, как можно судить по материалам конференции [3], состоявшейся в 2001 г. в г. Лувене (Бельгия) ("LINC-2001"), явно выраженный "аннотационный крен". В фокусе внимания находится формат аннотации (разметки) корпусов, методы борьбы с ошибками, даже психологическая реальность аннотированного корпуса (доклад А. Абейе и др.), а ведь проблема того, *что* именно за тексты должны быть представлены в корпусе и какие типы корпусов могут (должны?) различаться по этому параметру, не менее важна.

Основная задача заключается в репрезентативности корпуса. Общепризнанно, что письменный и устный язык функционирует во многих регистрах (жанрах и пр. — термины в разных традициях различны) — от литературы религиозной до технической, от публичной политической речи до неформальной беседы. Коль скоро эти регистры различаются на многих языковых уровнях, то и в корпусе должны быть представлены все или большинство из них. Как гласит нынешнее присловье, *в*

Интернете можно найти только то, что туда положили; то же относится и к корпусу. Решение этой задачи, несомненно, связано с охватом текстов. Принцип "чем больше, тем лучше" здесь работает; частотность и релевантность любого лингвистического явления проще проследить на объёме в 100 млн словоупотреблений, чем в 100 тыс. Такая задача стояла и перед создателями первых корпусов, в том числе Уппсальского, в который включены, наряду с художественными произведениями, публикации из журналов и газет.

Задача фиксации устной речи в корпусе вполне решаема, и зарубежный опыт (прежде всего немецкий) это показывает, но это задача для *отдельного корпуса;* необходимость транскрипции и желательность аудио- или видеодублирования неизбежно сообщают устному корпусу *иной формат,* а известные языковые особенности разговорной речи ставят её отдельно от письменной. Если мы вынесем устную речь за скобки, то мы должны вычленить из всего "моря" письменных русских текстов некоторые подмножества в зависимости от наших задач.

1.1. Полный корпус

Для некоторых задач, кажется, никакое *собственное* подмножество не может быть решением. Например, лингвист хочет выяснить, употреблялась ли вообще в XIX столетии такая-то лексема и в каком значении. Схожие задачи могут стоять и перед этимологом; на Западе есть давняя и прочная традиция указывать точную "дату рождения" слова в исторических словарях. Решение тут одно — *полный* корпус языка, в него входят *все* тексты (хотя бы печатные) на данном языке; единственным параметром его может быть время.

Разумеется, что для литературных языков XX (и второй половины XIX) в. такой корпус нереален, и единственное, что возможно — это пропорциональное к нему приближение. Но для более ранних эпох подобный банк данных вполне представим; например, корпус русского языка XVIII в. или польского XVI в. Кстати, и тот и другой корпус фактически возникли во время подготовки соответствующих академических словарей: не было только компьютерного представления.

1.2. Культурно-репрезентативный корпус

Сколько бы ни декларировалась независимость лингвистики от эстетико-культурного подхода, *письменный* язык — это прежде всего язык *культуры.* От лингвистической работы часто требуется не языковой пример, а (сколь угодно субъективно понимаемый) *хороший* языковой пример, "освященный" авторитетом сочинителя; характерно, что такое требование возникает во всех лингвистических традициях. Поэтому *собрание культурно значимых текстов* на данном языке также представляет собой обладающий собственной ценностью источник для аннотированного корпуса. Примером может служить создаваемый в настоящее время Австрийский национальный корпус в Вене, куда включаются тексты на немецком языке, оказавшие то или иное влияние на культуру Австрии, в том числе и переводные тексты, например, Библия или романы Достоевского (в разных переводах). В практическом преломлении "культурная значимость", во-первых, означает, что текст является потенциальным источником расхожих цитат (что важно при оценке частотности того или иного языкового явления). Во-первых, текст признаётся принадлежащим данному слою, если он вносит какой-то вклад в историю русского языка (в том числе и интересен языковыми экспериментами). Таким образом, это не что иное, как антология: сюда попадут все русские прозаики[1] первого-второго ряда (в перспективе возможно также — и культурно значимые переводы, в том числе и Библия). Этот корпус можно условно назвать *культурно-репрезентативным.*

1.3. "Эталонный" корпус

Но культурная значимость не гарантирует "стандартности" авторского языка. От корпуса языка требуется соответствие узусу и языковой компетенции его образованного носителя. Между тем тексты таких авторов, как Зощенко, Платонов или даже Гоголь изобилуют фразами, которые не будут грамматически правильными с точки зрения современной авторам нормы. Далёк от воплощения нормы полный корпус или его пропорциональное приближение. Отсюда задача: "*стандартный*", "*эталонный*" корпус русского языка; языка лишённого по возможности сознательных стилевых и лексических экспериментов, тем не менее "гладкого" и "профессионального". Здесь мы еще ближе к предельно субъективной оценке критериев включения/невключения того или иного текста или автора в корпус. Сюда не попадут, конечно, ни Зощенко, ни Платонов, ни, может быть, Солженицын, однако "пойдут" такие писатели второго ряда, как Трифонов или Рыбаков, язык которых может считаться достаточно "нейтральным" и "правильным", и даже, возможно, некоторые представители массовой литературы (такие тексты, как детективы, любовные романы и проч.). Поэтому возможно и расширение "стандартного" корпуса за рамки художественной литературы, с включением, например, публицистики. Разумеется, отбор текстов — задача сугубо индивидуальная для каждого автора, а порой и для отдельных произведений.

Такой корпус и создаётся в настоящее время в рамках проекта корпуса "ЦЛД-МГУ".

Таковы три точки на "шкале корпусов", организованной по параметру "ширина охвата текстов". Другая шкала — шкала времени. Отдельные корпуса должны быть созданы для XIX и XX вв., в пределах XX в. может быть выделен *современный* корпус, начало которого можно отнести к середине 1960-х гг. Именно создание современного корпуса, включающего в себя прозу 1965–2000 гг., и является нашей текущей первоочередной задачей.

2. ИСТОЧНИКИ ТЕКСТОВ

Источники текстов для корпусов в настоящее время весьма обширны. Если в 1960-х гг. всякий текст приходилось ad hoc представлять в

[1] Поэзия и драматургия — опять-таки иной вид речи; поэтический корпус, если будет создан (а это крайне интересная и полезная для стиховедения задача), будет корпусом совсем другого формата.

электронном виде, то в настоящее время практически все жанры русского письменного языка обширно представлены в Интернете. Особенностью Рунета являются *библиотеки текстов*, содержащие огромные коллекции как художественных, так и технических, правовых, публицистических и проч. произведений. На европейских и американских Интернет-сайтах не так легко найти коллекцию художественных текстов; зачастую это запрещено авторскими правами. У нас же имеются еженедельно обновляющаяся библиотека Максима Мошкова (www.lib.ru), коллекции "Общий текст" (www.textshare.da.ru), "Русский текст" (www.russiantext.com), а также такие сайты, как www.klassika.ru, www.divanchik.net и множество других. Как правило, тексты сканируют "на общественных началах" пользователи Интернета и присылают их администраторам библиотек. Таким образом, в распоряжении создателя корпусов находятся целые массивы художественных текстов в неразмеченном электронном виде — от Пушкина и Достоевского до Пелевина и Акунина, — из которого только надлежит выбрать нужное. Немало интересного для составителя корпуса содержат и сайты литературных журналов, в том числе таких, как "Новый мир", "Знамя" ("Журнальный зал" на www.infoart.ru). Не художественные тексты, прежде всего публицистика, новости, в меньшей степени — научные статьи также обильно представлены в русском секторе Всемирной сети. Это и официальные сайты газет и политических журналов, и "новостные" серверы, и различные образовательные ресурсы.

Основными сложностями, с которыми приходится сталкиваться при превращении "текстов для читателя" в "тексты для исследователя", является отсутствие единого стандарта подачи текста (в том числе даже в пределах одного ресурса: так, на www.lib.ru тире передаётся то дефисом в пробелах, то двумя дефисами, абзацный отступ — то тремя неразрывными пробелами, то пятью), большое подчас количество опечаток (для отсканированных текстов — ошибок распознавания), сохранение в тексте переносов, номеров страниц, иногда даже оформление строчек как абзацев (так называемые "жёсткие концы") и проч. Имеются методы полуавтоматического устранения таких неудобств, но опечатки — наиболее неприятные и трудноустранимые погрешности, серьёзно, как выяснилось, затрудняющие процесс "корпоризации" библиотечных текстов. В текстах Интернет-новостей, а также во многих других местах "неофициального" Интернета встречаются, наряду с опечатками, и просто орфографические и пунктуационные ошибки.

3. КОНЦЕПЦИЯ И ПРАКТИКА МОРФОЛОГИЧЕСКОЙ РАЗМЕТКИ

3.1. Морфология и/или синтаксис в корпусе

Следующим этапом технологического процесса на пути от неразмеченного текста к корпусу в Интернете является *разметка*. Напомним, что под разметкой (англ. annotation) понимается содержащаяся в корпусе лингвистическая или иная информация, приписанная тем или иным отрезкам текста (так, информация о времени создания и жанре приписывается целому тексту, информация о синтаксической структуре — предложению, информация о лексеме и грамматических характеристиках — слову), а равно и процесс, добавляющий к исходному тексту эту информацию.

В этой связи остановимся на одной важной для теории создания корпусов проблеме. Как в отечественной, так и в зарубежной традиции основное внимание уделяется синтаксической разметке. В полной мере этот относится и к "корпусу ИППИ". Морфологический и синтаксический анализы ("парсинг") в нём осуществляются при помощи механизмов, используемых в машинном переводе, что в значительной степени определяет принципы разметки на морфологическом уровне: например, последовательно различаются омонимичные наречия/союзы, такие, как *едва, иначе, отчего, когда, пока, ровно, точно, словно;* наречия/деепричастия, такие, как *стоя, сидя, лёжа,* что может быть проинтерпретировано как совмещение синтаксических и морфологических параметров. Такое деление имеется в традиционной русистике (и в "Грамматическом словаре" А. А. Зализняка); но объединение видовых пар в одну лексему (*пришла* анализируется как форма от *приходить*) явно продиктовано задачами машинного перевода.

Кроме того, никогда еще, насколько нам известно, не формулировалась цель создания корпуса, отражающего специально морфологию — уровень, стоящий ближе, чем синтаксический, к наивному восприятию носителя языка. А это отдельная и весьма полезная задача. Действительно, с точки зрения русской морфологии, не существует лексемы *сидя* и не различаются наречие и союз *пока;* реализация этих различий на собственно морфологическом уровне разметки есть не что иное, как "протаскивание" в морфологию синтаксических отношений. Синкретизм краткой формы прилагательного и наречия на *-о/-е* также близок к регулярности, и здесь омонимии можно не выделять (как и в немецком, где существует лишь ограниченное количество "чистых" наречий, а остальные омонимичны прилагательным).

Такой подход, помимо научной релевантности, имеет и ценность практическую. Он позволяет разметить большие объёмы текстов почти автоматически, значительно сократив при этом ручной этап разметки корпусов: не приходится уже каждый раз выбирать между союзом и частицей *и* или тремя вариантами *раньше* — сравнительной степенью от *рано*, сравнительной степенью от *ранний* и отдельным наречием.

Между тем задача *синтаксической разметки* может на данном этапе и не ставиться. Известно, что различные теории синтаксиса не сводятся к общему знаменателю, в отличие от морфологических (так, большинство зарубежных корпусов используют генеративистский подход, корпус ИППИ — модель синтаксических зависимостей, восходящую к соответствующему уровню модели "Смысл⇔Текст"). Разумеется, и морфологическая разметка не может не потребовать от создателя корпуса определённой концепции; но поиск определённых синтаксических конструкций вряд ли возможен без подробного предварительного изучения используемого в данном корпусе формализма.

Пока мы используем из синтаксического уровня разметки только ярлык (тег) <s>, разделяющий предложения (при этом сохраняется HTML-ярлык

<p> для абзаца). Заметим, что в нашем корпусе будет вообще сохраняться HTML-разметка исходных "библиотечных" файлов (выделение заголовков, шрифтовые выделения в тексте и проч.). Для поиска необходимо добавить также разметку по следующим полям: "автор текста", "название текста", "дата создания текста" (в файлах, размещенных на сайте "Библиотека Мошкова", сохраняется авторская датировка), "жанр".

3.2. Морфологическая разметка. Этап первый: анализ (парсинг)

Морфологическая разметка осуществляется при помощи программы (парсера) MYSTEM, написанного программистами компании "Яндекс" И. В. Сегаловичем и В. А. Титовым. MYSTEM написан в среде Linux, имеется возможность работы в среде Microsoft Windows. В основе программы — алгоритм "Грамматического словаря русского языка" А. А. Зализняка. На входе имеем файл в формате HTML или TXT (кодировка Windows), на выходе — файл, где после каждой словоформы в фигурных скобках через знак | указаны возможные варианты разбора:

{лексема$_1$ = грамматические признаки лексемы$_1$ = грамматические признаки словоформы$_1$ | лексема$_1$ = грамматические признаки лексемы$_1$ = грамматические признаки словоформы$_2$ | лексема$_2$ = ...}

Для словоформ, отсутствующих в словаре, указываются со знаком "?" гипотезы.

Для примера проведём пропущенный через парсер текст (символы исходного текста для удобства чтения выделены полужирным шрифтом):

Я сидел на барском сиденье, дышал горячим ветром, бившим в лицо, ощущая в то же время не истребимую никакими сквозняками пыль и легкий запах духов — катафалк с хорошей скоростью мчался по шоссе на юг. (Ю. Трифонов)

{\s} **Я**{я=S, сред, неод=им, ед | я=S, сред, неод=им, мн | я=S, сред, неод=ро
д, ед | я=S, сред, неод=род, мн | я=S, сред, неод=дат, ед | я=S, сред, неод=д
ат, мн | я=S, сред, неод=вин, ед | я=S, сред, неод=вин, мн | я=S, сред, неод=
твор, ед | я=S, сред, неод=твор, мн | я=S, сред, неод=пр, ед | я=S, сред, нео
д=пр, мн | я=S, ед, од=им, жен | я=S, ед, од=им, муж} **си-**
дел{сидеть=V, несов=прош, ед, изъяв, муж} **на**{на=PART= | на=PR=}
барском{барский=A=пр, ед, муж | барский=A=пр, ед, сред} **сиде-**
нье{сиденье=S, сред, неод=им, ед | сиденье=S, сред, неод=вин, ед | сиден
ье=S, сред, неод=пр, ед}, **дышал**{дышать=V, несов=прош, ед, изъяв, муж}
горя-
чим{горячий=A=дат, мн | горячий=A=твор, ед, муж | горячий=A=твор, ед, с
ред | горячее=S, ед, сред, неод=твор | горячить=V, несов=непрош, ед, при
ч, кр, муж, страд | горячить=V, несов=непрош, мн, изъяв, 1-л} **вет-**
ром{ветер=S, муж, неод=твор, ед}, **бив-**
шим{бить=V, несов=прош, дат, мн, прич | бить=V, несов=прош, твор, ед, пр
ич, муж | бить=V, несов=прош, твор, ед, прич, сред} **в**{в=PR=} **ли-**
цо{лицо=S, сред, неод=им, ед | лицо=S, сред, неод=вин, ед | лицо=S, сред,
од=им, ед | лицо=S, сред, од=вин, ед}, **ощу-**
щая{ощущать=V=непрош, деепр, несов} **в** {в=PR=}
то{то=CONJ= | тот=A=им, ед, сред | тот=A=вин, ед, сред | то=S, ед, сред, не
од=им | то=S, ед, сред, неод=вин}
же{же=PART=|же=S, сред, неод=им, ед | же=S, сред, неод=им, мн | же=S, сре
д, неод=род, ед | же=S, сред, неод=род, мн | же=S, сред, неод=дат, ед | же=S
, сред, неод=дат, мн | же=S, сред, неод=вин, ед | же=S, сред, неод=вин, мн |
же=S, сред, неод=твор, ед | же=S, сред, неод=твор, мн | же=S, сред, неод=п
р, ед | же=S, сред, неод=пр, мн | же=CONJ=} **вре-**
мя{время=S, сред, неод=им, ед | время=S, сред, неод=вин, ед}
не{не=PART=} **истребимую**{истребимый=A=вин, ед, жен} **никаки-**
ми{никакой=A=твор, мн} **сквозняками**{сквозняк=S, муж, неод=твор, мн}
пыль{пыль=S, ед, жен, неод=им | пыль=S, ед, жен, неод=вин}
и{и=PART= | и=S, сред, неод=им, ед | и=S, сред, неод=им, мн | и=S, сред, нео
д=род, ед | и=S, сред, неод=род, мн | и=S, сред, неод=дат, ед | и=S, сред, не
од=дат, мн | и=S, сред, неод=вин, ед | и=S, сред, неод=вин, мн | и=S, сред, н
еод=твор, ед | и=S, сред, неод=твор, мн | и=S, сред, неод=пр, ед | и=S, сред
, неод=пр, мн | и=INTJ= | и=CONJ=} **лег-**
кий{легкий=A=им, ед, муж | легкий=A=вин, ед, муж, неод}
за-
пах{запах=S, муж, неод=им, ед | запах=S, муж, неод=вин, ед | запах=S, муж
, неод=им, ед | запах=S, муж, неод=вин, ед | запахнуть=V, сов=прош, ед, из
ъяв, муж} **ду-**
хов{духов=A=им, ед, муж | духов=A=вин, ед, муж, неод | дух=S, муж, неод=р
од, мн | дух=S, муж, од=род, мн | дух=S, муж, од=вин, мн | духи=S, мн, муж,

неод=род} - - **ката-**
фалк{катафалк=S, муж, неод=им, ед | катафалк=S, муж, неод=вин, ед}
с{с=PR=} **хорошей**{хорошеть=V, несов=ед, пов, 2-
л | хороший=A=род, ед, жен | хороший=A=дат, ед, жен | хороший=A=твор, ед,
жен | хороший=A=пр, ед, жен} **скоро-**
стью{скорость=S, жен, неод=твор, ед} **мчал-**
ся{мчаться=V, несов=прош, ед, изъяв, муж | мчать=V, несов=прош, ед, изъ
яв, муж, страд} **по**{по=PR=} **шос-**
се{шоссе=S, сред, неод=им, ед | шоссе=S, сред, неод=им, мн | шоссе=S, сре
д, неод=род, ед | шоссе=S, сред, неод=род, мн | шоссе=S, сред, неод=дат, е
д | шоссе=S, сред, неод=дат, мн | шоссе=S, сред, неод=вин, ед | шоссе=S, ср
ед, неод=вин, мн | шоссе=S, сред, неод=твор, ед | шоссе=S, сред, неод=тво
р, мн | шоссе=S, сред, неод=пр, ед | шоссе=S, сред, неод=пр, мн}
на{на=PART= | на=PR=}
юг{юг=S, муж, неод=им, ед | юг=S, муж, неод=вин, ед}.

На данном этапе имеются две сложности, обе связанные со словником Грамматического словаря и на первый взгляд противоречащие друг другу. Во-первых, речь идёт об *ограниченности* словника словаря Зализняка, в котором отсутствуют имена собственные, некоторые неологизмы последнего времени, сравнительные формы вроде *постарше,* наречия вида *по-детски,* многие сложные слова, пишущиеся через дефис, многие наречия на *-о* и *-е* (последняя задача не снимается введением синкретического класса "наречие/краткая форма прилагательного"). Проблема расширения словаря в настоящее время решается; так, в отдельные "словарные статьи" выделены употребляющиеся только через дефис компоненты сложных слов, такие, как *англо-, темно-, русско-, человеко-, машинно-* и др., что позволит не добавлять в словарь все слова, образованные при помощи этих чрезвычайно продуктивных компонентов.

Во-вторых, множество порождаемых словоформ *излишне широко* с точки зрения вероятности встречаемости данных форм в тексте. Последнее обстоятельство сильно затрудняет снятие омонимии. Например, цепочка букв *их* получает разбор "междометие", цепочка *он* — разбор "существительное", т. е. название буквы О (а ведь это не один, а целых 12 разборов, все падежно-числовые формы несклоняемого слова!), цепочка *полей* интерпретируется, помимо родительного падежа множественного числа от *поле* и императива от *полить,* еще и как — фактически невозможная — сравнительная форма от *полый* (пример С. А. Крылова). Добавим сюда и заметное в нашем примере разграничение мужской и женской "форм" для слов *я* или *ты.* Разумеется, все такие случаи "нежелательной омонимии" предусмотреть и отсеять невозможно, тем не менее подобные разборы наиболее частотных словоформ следует "отсекать" автоматически.

3.3. Морфологическая разметка. Этап второй: фильтрирование

Проанализированный текст проходит через фильтр GRAMBAT на языке "Perl" (автор А. Е. Поляков), который удаляет варианты разбора, вероятность которых близка к нулю, но которые тем не менее порождаются при помощи алгоритма словаря Зализняка (например, анализ цепочки символов *при* как императива от *перть* или форм от существительного *пря,* анализ *и* как названия буквы И), объединяет некоторые омонимичные формы вроде вышеуказанных наречий-союзов, а также помечает знаком "?" варианты разбора, не соответствующие синтаксическому окружению (например, *шоссе* как именительный падеж после предлога *по*). GRAMBAT — это "пакетный файл" (batch file), запускающий последовательно MYSTEM и программу фильтрования. Промежуточные результаты тем не менее сохраняются для дальнейшей отладки.

Именно здесь, на этапе фильтрования, вносятся элементы принимаемой нами морфологической концепции, которая сводится к минимизации межчастеречной омонимии слов и словоформ. Например, мы "не признаём" наречий *утром, вечером, порой,* не считаем, что у "тривиально" субстантивируемых прилагательных *молодая, старший, ссыльнокаторжная* есть разбор "существительное". Наиболее "революционный" шаг — введение "синкретических" категорий: "наречие/краткая форма прилагательного" (большинство наречий на *-о* и *-е*) или "наречие/союз" (см. выше).

Вот как выглядит разметка на этом этапе:

<s>**Я**{я=S, ед, од=им} **си-**
дел{сидеть=V, несов=изъяв, прош, ед, муж} **на**{на=PR | на=PART}
барском{барский=A=ед, муж, пр | барский=A=ед, сред, пр} **сиде-**
нье{сиденье=S, сред, неод=ед, им? | сиденье=S, сред, неод-ед, вин | сиде
нье=S, сред, неод=ед, пр}, **ды-**
шал{дышать=V, несов=изъяв, прош, ед, муж} **горя-**
чим{горячий=A=мн, дат | горячий=A=ед, муж, твор | горячий=A=ед, сред, т
вор | горячить=V, несов=изъяв, непрош, мн, 1-л} **вет-**
ром{ветер=S, муж, неод=ед, твор}, **бив-**
шим{бить=V, несов=прич, прош, мн, дат | бить=V, несов=прич, прош, ед, му
ж, твор | бить=V, несов=прич, прош, ед, сред, твор} **в**{в=PR} **ли-**
цо{лицо=S, сред, неод=ед, им? | лицо=S, сред, неод=ед, вин | лицо=S, сред
, од=ед, им? | лицо=S, сред, од=ед, вин}, **ощу-**

щая{ощущать=V=несов, деепр, непрош} **в**{в=PR}
то{то=CONJ | тот=A=ед, сред, им? | тот=A=ед, сред, вин | то=S, сред, неод, ед=им? | то=S, сред, неод, ед=вин} **же**{же=PART} **вре-**
мя{время=S, сред, неод=ед, им | время=S, сред, неод=ед, вин}
не{не=PART} **истребимую**{истребимый=A=ед, жен, вин} **никаки-**
ми{никакой=A=мн, твор} **сквозняками**{сквозняк=S, муж, неод=мн, твор}
пыль{пыль=S, жен, неод, ед=им | пыль=S, жен, неод, ед=вин} **и**{и=CONJ}
легкий{легкий=A=ед, муж, им | легкий=A=ед, муж, вин, неод}
за-
пах{запах=S, муж, неод=ед, им | запах=S, муж, неод=ед, вин | запах=S, муж, неод=ед, им | запах=S, муж, неод=ед, вин | запахнуть=V, сов=изъяв, прош, ед, муж} **ду-**
хов{духов=A=ед, муж, им | духов=A=ед, муж, вин, неод | дух=S, муж, неод=мн, род | дух=S, муж, од=мн, род | дух=S, муж, од=мн, вин | духи=S, муж, неод, мн=род} - - **ката-**
фалк{катафалк=S, муж, неод=ед, им | катафалк=S, муж, неод=ед, вин}
с{с=PR=} **хорошей**{хорошеть=V, несов=пов, ед, 2-
л | хороший=A=ед, жен, род | хороший=A=ед, жен, дат? | хороший=A=ед, жен, твор | хороший=A=ед, жен, пр?} **скоро-**
стью{скорость=S, жен, неод=ед, твор} **мчал-**
ся{мчаться=V, несов=изъяв, прош, ед, муж | мчать=V, несов=изъяв, прош, ед, муж, страд} **по**{по=PR} **шос-**
се{шоссе=S, сред, неод=ед, им? | шоссе=S, сред, неод=мн, им? | шоссе=S, сред, неод=ед, род? | шоссе=S, сред, неод=мн, род? | шоссе=S, сред, неод=ед, дат | шоссе=S, сред, неод=мн, дат | шоссе=S, сред, неод=ед, вин? | шоссе=S, сред, неод=мн, вин? | шоссе=S, сред, неод=ед, твор? | шоссе=S, сред, неод=мн, твор? | шоссе=S, сред, неод=ед, пр? | шоссе=S, сред, неод=мн, пр? }
на{на=PR|на PART}
юг{юг=S, муж, неод=ед, им? | юг=S, муж, неод=ед, вин}.

3.4. Морфологическая разметка. Этап третий: снятие омонимии

Наконец, текст проходит процесс снятия омонимии, осуществляемый вручную. При этом используется программа GRAMEDIT, написанная А. Е. Поляковым на языке макросов Microsoft Word (это подключаемый к программе Microsoft Word "шаблон" формата *.dot). Проинструктированный оператор проходит все слова с числом разборов, не равным одному (или с единственным разбором-гипотезой), выбирая нужный. При необходимости оператор редактирует вариант или вводит новый. Разметка всех прочих слов при этом скрыта для удобства просмотра контекста. Предоставлены возможности "отката" (отмены предыдущего исправления) и глобальной замены по всему тексту. Во всех случаях, когда оператор не уверен в правильности выбора того или иного анализа для некоторого слова, он должен это слово пропустить и предоставить окончательный выбор одному из руководителей проекта.

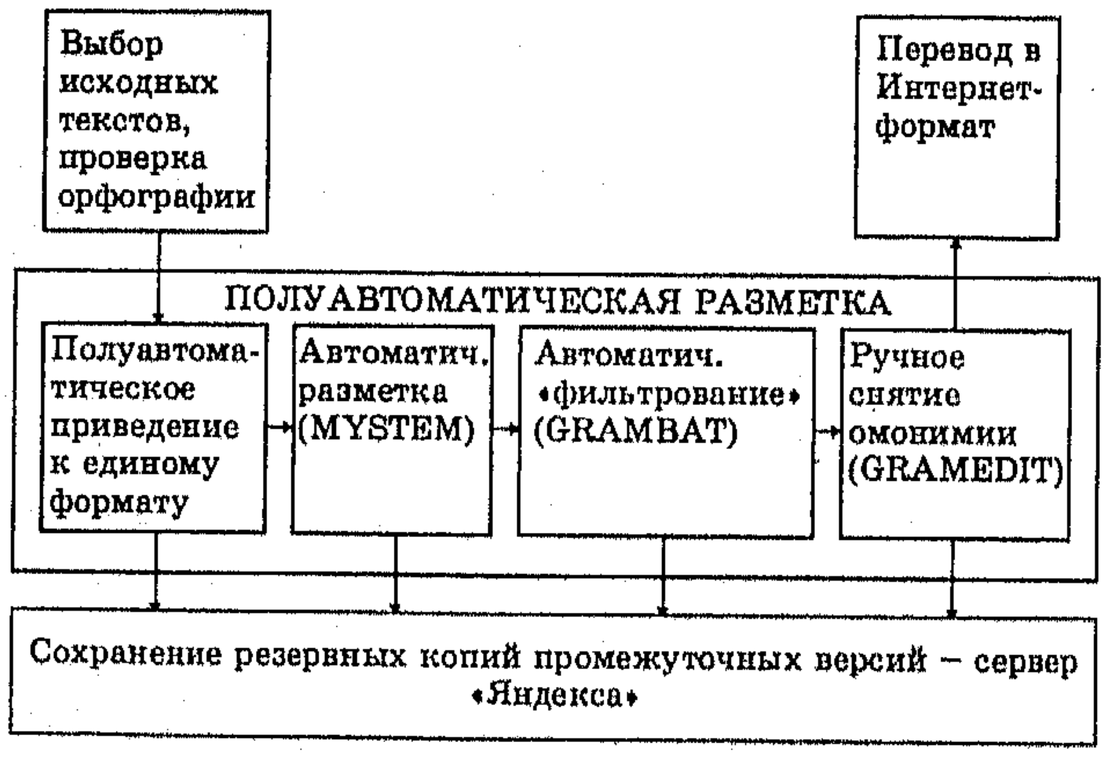
Пример из Ю. Трифонова после снятия омонимии принимает следующий вид:

<s>**Я**{я=S, ед, од=им} **сидел**{сидеть=V, несов=изъяв, прош, ед, муж}
на{на=PR}
барском{барский=A=ед, сред, пр} **сиде-**
нье{сиденье=S, сред, неод=ед, пр}, **ды-**
шал{дышать=V, несов=изъяв, прош, ед, муж} **горя-**
чим{горячий=A=ед, муж, твор} **ветром**{ветер=S, муж, неод=ед, твор},
бившим{бить=V, несов=прич, прош, ед, муж, твор} **в**{в=PR} **ли-**
цо{лицо=S, сред, неод=ед, вин}, **ощу-**
щая{ощущать=V=несов, деепр, непрош} **в**{в=PR}
то{тот=A=ед, сред, вин} **же**{же=PART} **вре-**
мя{время=S, сред, неод=ед, вин} **не**{не=PART} **истреби-**
мую{истребимый=A=ед, жен, вин} **никакими**{никакой=A=мн, твор}
сквозняками{сквозняк=S, муж, неод-мн, твор}
пыль{пыль=S, жен, неод, ед=вин} **и**{и=CONJ} **лег-**
кий{легкий=A=ед, муж, вин, неод}
запах{запах=S, муж, неод=ед, вин} **духов**{духи=S, муж, неод, мн=род} -
- **катафалк**{катафалк=S, муж, неод=ед, им} **с**{с=PR} **хоро-**
шей{хороший=A=ед, жен, твор} **скоро-**
стью{скорость=S, жен, неод=ед, твор} **мчал-**
ся{мчаться=V, несов=изъяв, прош, ед, муж} **по**{по=PR} **шос-**
се{шоссе=S, сред, неод=ед, дат} **на**{на=PR}
юг{юг=S, муж, неод=ед, вин}.

В настоящее время осуществляется (силами участников проекта, сотрудников ОЛИ ВИНИТИ и студентов отделения теоретической и прикладной лингвистики филфака МГУ) полуавтоматическое снятие омонимии. На следующем этапе размеченный таким образом текст проходит обработку для размещения его в составе Интернет-корпуса.

Для отслеживания ошибок и восстановления первоначального варианта разметки в случае необходимости все этапы разметки текста должны сохраняться на особом сервере в компании "Яндекс", с указанием номера "версии" того или иного текста.

Таким образом, весь процесс переработки исходного неразмеченного "библиотечного" текста в размещаемый в корпусе размеченный текст выглядит следующим образом:



4. ИНТЕРФЕЙС В СЕТИ

Мы предполагаем, что пользователю через сетевой интерфейс будут предоставлены следующие возможные операции с текстом — просмотр (как в "библиотеке") и поиск (с последующим просмотром). При просмотре должно быть предусмотрено переключение между размеченным и неразмеченным представлением текста (как это имеет место в Британском национальном корпусе). Другой крайне интересной возможностью является представление текста в виде последовательности активных ссылок. "Щелчок" на каждом слове активизирует открытие нового окна с указанием лексемы, словоформы, возможно также — словарной статьи (статей) и проч. По такому принципу организован уже упоминавшийся корпус латинских текстов "Персей" (www.perseus.tufts.edu). В нашем корпусе, по-видимому, будет реализован формат "переключение". Пользователь также должен иметь доступ к документации корпуса ("руководству по эксплуатации", списку авторов и произведений и др.).

Поиск должен быть возможен по следующим параметрам (в скобках указаны примеры):

1) конкретная словоформа (*большого*);

2) лексема (*большой*) — выдаёт *большому, больших...*;

3) морфологические параметры ("все прилагательные в творительном падеже единственного числа мужского рода"; "все существительные pluralia tantum");

4) линейная позиция относительно другой лексемы/словоформы, абсолютная или относительная ("все контексты, содержащие слова, начинающиеся на *рас-*, непосредственно после лексемы *слегка*");

5) линейная позиция в предложении ("все вхождения *и* в начале предложения");

6) число вхождений в предложении/абзаце;

7) дата и тип текста ("примеры из художественной литературы второй половины XIX в."; "примеры из произведений Достоевского");

а также по комбинациям данных параметров (например, "сочетание предлога *в* с любым словом во втором предложном падеже"), в том числе и с использованием логических символов И, ИЛИ и НЕ ("...во втором предложном или предложном падеже", "... во втором предложном, но не в предложном падеже" и т. д.). Должна быть предусмотрена возможность использования во всех полях специальных знаков * и ? (любая последовательность символов, любой символ), возможность выбирать формат выдачи (длина контекста в предложениях/абзацах, количество выдаваемых контекстов на одной странице и пр.), а также возможность вторичного поиска в найденном и сохранения результатов поиска. Оболочка для сетевого интерфейса нашего корпуса пишется программистами компании "Яндекс". Не исключено создание многоязычного интерфейса на основных европейских языках.

В отечественном Интернете уже имеется корпус с поисковой оболочкой такого типа (правда, без возможности поиска по синтаксическим параметрам и просмотра текста в режиме разметки) — это составленный одним из участников нашего проекта А. Е. Поляковым "Словарь языка Грибоедова" (www.inforeg.ru/concord/index.htm).

Таков в общих чертах проект русских корпусов в Интернете "ЦЛД-МГУ". Он предусматривает возможность постоянного пополнения после того, как в Сети будет размещён некоторый "стартовый" объём текстов. Надеемся, он может послужить достойным заполнением досадной лакуны и облегчит труды русистов в России и в мире.

СПИСОК ЛИТЕРАТУРЫ

1. Андрюшенко В. М. Концепция и архитектура машинного фонда русского языка.— М., 1989.

2. Богуславский И. М. и др. Аннотированный корпус русских текстов: концепция: инструменты разметки, типы информации // Тр. Междунар. семинара по компьютерной лингвистике и её приложениям "Диалог-2000".— Протвино, 2000.

3. Empirical Methods in the new Millenium: Linguistically interpreted Corpora. Programme and Abstracts // Электронная публикация по адресу: http://wwwling.arts.kuleuven.ac.be/sle2001/empirical-programme.htm

Материал поступил в редакцию 05.07.02.