К СОЗДАНИЮ ПРЕДСТАВИТЕЛЬНОГО КОРПУСА СОВРЕМЕННОГО РУССКОГО ЯЗЫКА

С. А. Шаров, РосНИИ Искусственного Интеллекта,
125190, Москва, а/я 85, sharoff@aha.ru

В статье вводятся основные понятия, необходимые для создания представительного корпуса русского языка, описываются современное состояние русских корпусов, описываются принципы построения предлагаемого представительного корпуса и проблемы, возникающие при его разработке. Корпус объемом примерно в 100 млн. слов должен обеспечить пропорциональное покрытие всех основных речевых жанров, что позволит получать статистически достоверную об использовании слов и грамматических конструкций в современном русском языке. В статье рассматривается три вида проблем: получение исходных текстов, разрешение неоднозначности при разметке и построение формального языка запросов к корпусу.

Введение

С точки зрения корпусной лингвистики русский язык является одним из немногих мировых языков, не имеющих представительного корпуса, отражающего современное состояние и использование языка. Вместе с тем необходимость в создании такого корпуса ясно осознается в лингвистическом сообществе в России и за ее пределами.

В пределах СССР в 1970-е годы был создан частотный словарь русского языка (Засорина, 1977) на основе корпуса текстов объемом 1 миллион слов, включавшего примерно в равной пропорции общественно-политические тексты, художественную литературу, научные и научно-популярные тексты из разных областей и драматургию (тексты последней области были предназначены для приближенного отражения параметров устной речи). Хотя корпус был довольно аккуратно проработан (включая ручное внесение лемматизации и частеречной аннотации), как корпус он недоступен.

Самый известный представительный корпус русского языка был создан в Финляндии в Университете Уппсалы (УК). Уппсальский корпус объемом примерно в 1 миллион слов состоит из художественной литературы и информативных текстов, примерно в равной пропорции, см. (Lönngren, 1993). Художественная литература взята за период 1960-1988 (40 писателей, фрагменты 300 текстов), информативная проза взята за период 1985-1988. С современной точки зрения корпус слишком мал и ограничен в отражении речевых жанров. В нем также отсутствуют лемматизация и частеречная аннотация, что существенно ограничивает возможности его использования для поиска различных форм одного слова.

Известная попытка создания представительного корпуса была предпринята А.П. Ершовым и В.М. Андрющенко в первой половине 80-х под названием Машинного Фонда русского языка, см. (Ершов, 1981), (Андрющенко, 1989). Цели этого проекта были близки целям создания Британского Национального Корпуса (БНК), который начал создаваться несколькими годами позднее. К сожалению, этот проект не закончился созданием собственно корпуса, хотя было собрано много текстов разного типа. Существуют также многочисленные коллекции русских текстов, но они не сбалансированы, обычно ограничены областью художественной литературы и не имеют адекватных механизмов поиска.

В связи с отсутствием лучшего механизма исследования русского языка на основе корпуса в настоящее время используют поиск в Интернете. Например, с помощью поисковой машины можно провести статистический анализ того, как часто встречаются словосочетания *истинное наслаждение* и *истинное удовольствие*, см. (Levontina, Zalizniak, 2001), (Sharoff, 2002). Интернет можно рассматривать как самый большой корпус русского языка: количество текстов на русском языке в Интернет можно оценить объемом 250 млрд. слов (1.5 терабайта уникальных текстов, проиндексированных Яндексом), что намного больше любого возможного корпуса. Но

Интернет не является корпусом, поскольку тексты представлены в Интернет случайным образом: это множество зависит от предпочтений и интересов очень специфичной группы носителей русского языка, а именно активных пользователей Интернета, например, поиск употреблений словосочетаний *истинное наслаждение* и *истинное удовольствие* приводит, главным образом, к порнографическим сайтам, что вряд ли соответствует общеязыковым контекстам использования этих словосочетаний. Кроме того, поисковые машины позволяют найти все формы слова (хотя они и не используют лемматизацию в чистом виде), но лингвистически интересные запросы сложно, а часто и невозможно сформулировать в поисковой машине, особенно это касается грамматических признаков. Например, если нас интересует критика чего бывает в русском языке, мы не можем задать вопрос на поиск всех существительных в родительном падеже, следующих за словом *критика,* которая, помимо всего прочего, омонимична слову *критик*. Другие возможные примеры касются поиска употреблений второго предложного падежа или возможности комбинации некоторого глагола с существительными в дательном падеже.

Есть и еще одна проблема при поиске. Выделение лемм поисковыми машинами призвано обеспечить потребности обычного пользователя, которые могут отличаться от потребностей лингвиста. Например, нормальные пользователи, заинтересованные в доступе к информации, не обращают внимания на вид глаголов, используемых при поиске и это совершенно разумно. С точки зрения информационного поиска в ответ на запрос *Объявлял ли Путин о поддержке Хусейна* пользователь хочет получить страницу с текстом *В своей речи Путин объявил о поддержке Хусейна*. Поисковые машины следуют нуждам таких пользователей и неявным образом объединяют в запросе видовые пары (в идеале они должны также объединять *поддержка* и *поддерживать*). Но с точки зрения лингвистического поиска неразумно включать этот механизм без необходимости, поскольку глаголы противоположного вида могут отличаться своей семантикой. Наконец, поисковые машины активно используют поиск в списках ключевых слов и заголовках документов, что мешает работе лингвиста.

С точки зрения результатов поиска по запросу невозможно оценить их полноту и представительность, поскольку они зависят от неизвестных параметров: какие тексты не доступны на Интернет, или какие тексты, выложенные на Интернет, не были проиндексированы поисковой машиной, использованной в запросе. Также результаты запроса не могут быть представлены в лингвистически релевантном виде: поисковая машина предъявляет сами документы, отсортированные в соответствии с их релевантностью для инфомационного поиска. Наконец, Интернет является динамичной системой, поэтому нет возможности сравнивать результаты поисков, сделанных в разное время.

Цели создания корпуса

Целью данной работы является создание представительного корпуса русского языка, подобного БНК. Создаваемый большой корпус русского языка (БОКР) объемом примерно в 100 млн слов должен обеспечить пропорциональное покрытие всех основных речевых жанров, что позволит получать статистически достоверную об использовании слов и грамматических конструкций в современном русском языке. Корпус предполагает проведение лемматизации и частеречной аннотации, а также именных и предложных групп: синтаксическая разметка в полном виде отсутствует по причине ненадежности алгоритмов синтаксического разбора, поэтому синтаксические признаки используются, в первую очередь, для снятия неоднозначности омографов и разрешения падежной неоднозначности. Для корпуса будет также создан язык запросов, учитывающий особенности русской морфологии и синтаксиса. Созданный в результате этой работы корпус будет свободно доступен через Интернет для исследования современного русского языка.

Необходимость создания большого корпуса определяется тем, что язык со статистической точки зрения представляет собой большое количество редких событий, например, русское слово *ножницы* во всех своих формах встречается реже чем 10 раз на миллион слов, т.е. с вероятностью меньшей чем 10^{-5}. Поэтому для построения достоверной модели языка корпус должен быть достаточно большим, например, включать *ножницы* не менее 100 раз, чтобы иметь возможность описать это слово, включая *гидравлические* или *маникюрные ножницы*, которые имеют разные контексты использования, не считая *ножниц цен, доверия* или *"сломанных ножниц экономического образования"*. Корпус же небольшого размера, например, 1 млн. слов, подобно УК и корпусу словаря (Засорина, 1977), может случайным образом существенно исказить частоту и основные контексты словоупотребления. По этой причине с ростом вычислительных мощностей увеличивались и размеры корпусов. Первый представительный корпус, Брауновский Корпус (БК, Brown Corpus), имел размер 1 млн. слов (УК и корпус словаря (Засорина, 1977) были созданы по его модели: фрагменты приблизительно 500 текстов, каждый из которых имел объем около 2000 слов). БНК, созданный в начале 90-х, уже имел объем в 100 млн. слов и определил стандарт для создания представительных корпусов различных языков

С другой стороны, БОКР включает в себя "Русский Стандарт", корпус размером в 10 млн. слов, который отражает современное состояние литературного языка. Различие между корпусами связано с большим вниманием на большой размер корпуса, широкое покрытие видов использования языка и баланс жанров в случае БОКРа и на отбор текстов, представительных для литературного русского языка, и ручная проверка морфосинтаксической

разметки в Русском Стандарте. Ручное разрешение неоднозначности делает Русский Стандарт похожим на ядро БНК, корпус объемом 5 млн. слов, морфосинтаксическая разметка которого также была проверена вручную (Leech, 1997). При разработке корпусов предполагается, что Русский Стандарт служит основным источником данных для развития грамматик русского языка на основе корпуса, а БОКР служит вспомогательным источником редкой грамматической информации и основным источником информации о лексике.

В отношении пропорции жанров структура корпуса Русский Стандарт существенно отличается от принципов построения ядра БНК. Пропорция жанров текстов в ядре БНК следует пропорции жанров в полном корпусе, в то время как Русский Стандарт основан на художественной прозе. Разница в построении этих подкорпусов отражает различие в культурном статусе языка художественной прозы в британской и русской культурах. В русской культуре литературный язык считается авторитетным источником, который в сущности определяет норму языка, используемого его носителями. Подавляющее большинство реальных примеров, используемых в школьных учебниках русского языка, грамматиках и словарях, взято из художественных произведений (чаще всего из классической литературы). В британской же традиции многие примеры в учебниках и справочниках берутся из самых рахных источников. Более высокий культурный статус языка художественной литературы отражен и в существующих корпусах: и УК, и корпус, использованный в (Засорина, 1977) примерно на половину состоят из художественной прозы, что существенно больше пропорции художественной литературы в Брауновском Корпусе (25%) и БНК (17%). В БОКРе также предполагается иметь большее количество литературных текстов (около 30%). Параллельно с этим корпусом также в рамках программы РАН "Филология и информатика" в Санкт-Петербурге будет создаваться Национальный корпус русского языка XIX и первой половины XX вв. объемом также около 100 млн. слов. Целью этого проекта является сбор значимых для русского языка текстов в филологически достоверном виде, а также организация ввода текстов в компьютер с последующей вычиткой и их филологической обработкой (Вербицкая и др., 2003).

Проблемы создания корпуса и методы их решения

Можно выделить три вида таких проблем: получение исходных текстов, разрешение неоднозначностей разного уровня и построение формального языка запросов к корпусу.

Некоторые типы текстов легко доступны, например, художественная литература и газетно-журнальные тексты. Другие виды текстов, например, деловую или личную переписку гораздо сложнее получить и использовать при создании корпусов, хотя эти жанры весьма важны для построения представительного корпуса. Наконец, крайне трудно получить достаточное количество транскрибированной устной речи, особенно используемой в условиях, отличных от публичной речи, такой как доклады, прения и т.п. В то же время, именно последний тип речи представляет особый интерес. Он важен, в частности, для исследования грамматических и лексических отличий ее от публичной и письменной речи, поскольку это дает представление о реальном использовании современного русского языка подавляющим большинством его носителей. В связи с этим, при создании корпуса делается попытка при возможности сбалансировать количество текстов в корпусе в пользу "эфемерных жанров" (письменных и устных текстов, производимых непрофессиональными авторами), поскольку электронные версии художественной литературы и газетно-журнальных текстов слишком легко превзойдут их по объему.

Еще одна проблема с источниками связана с диахроническими параметрами выборки. Активный исторический процесс в СССР и России достаточно радикально менял русский язык на протяжении 20-го века. В связи с этим выбор хронологических рамок для создания корпуса существенно влияет на результаты. Например, в частотном словаре Засориной (1977) слова *советский, коммунистический, революция* и *товарищ*, входят в первую сотню русских слов, опережая многие служебные слова, такие как *ваш, лучше, здесь*. При построении частотного списка на основе газетно-журнальных текстов второй половины 1990-х эти же слова оказываются относительно редки (особенно *советский* и *товарищ,* чья частота в современном корпусе сравнима с частотой слов *греческий* или *сыр*).

В связи с тем, что политическая ситуация по-разному влияет на разные виды функциональных жанров, для описываемого корпуса выбор временного интервала для взятия соответствующих текстов варьируется. В частности, художественная литература берется начиная с 1970 года, научные тексты с 1980, общественно-политические тексты и "эфемерные жанры" с 1990 (это ограничение объяснимо и техническими причинами: более ранние тексты практически недоступны в электронном виде), а газеты и журналы с 1996.

Проблемы с разрешением неоднозначности связаны с тем, что несмотря на развитую морфологию многие словоформы в русском языке неоднозначны, например, *дорогой, были, три, уже*. В соответствии со списком Аношкиной в русском языке есть почти 20000 омографов. Помимо экзотических случаев (например, слова *для, задаст*), для которых выбор леммы статистически очевиден (вероятность деепричастия от *длить* и краткой формы от *задастый* близка к нулю), есть случаи регулярной омонимии, например, между кратким прилагательным и наречием (*абсурдно, трудно*), существительным и глаголом (*стали, пролив, полей*), существительным и прилагательным, включая местоименные формы (*дорогой, часовой, его*), между существительным и наречием

(*весной, пора*) и т.д. Особенно часты случаи омонимии между существительными внутри части речи, например, *поле* является формой *пол, поле* и *пола* (также *Поля* при написании с большой буквы).

В связи с тем, что морфологические признаки омографов различны, многие виды неоднозначности можно разрешить при помощи простых синтаксических алгоритмов построения именных и предложных групп, а также на основе статистических соображений удалением редких омографов, оставшихся *после* локального синтаксического анализа (*знаменитая гнома, данная метода* будут оставлены в женском роде). Оставшаяся неоднозначность должна быть сохранена в корпусе: без семантического и прагматического анализа полного текста невозможно решить, что имелось в виду в заголовке газетной статьи "*Не храните свои деньги в банке*".

Формат хранения корпуса основан на стандарте TEI (Sperberg-McQueen, Burnard, 2001), который используется для разметки большого количества корпусов, включая БНК. Морфологическая информация в нем хранится внутри специальных элементов (<ana>, analysis), неоднозначность представлена параллельными возможностями анализа:

```
<s id="kozlotur.1476">
  <w n="1">Мне<ana lemma="я" feats="МС,ед,1л: дт"/></w>
  <w n="2">было<ana lemma="быть" feats="Г,нс,нп,дст: ср,ед,прш"/></w>
 <phr type="ADV+ADV"> <w n="3">очень<ana lemma="очень" feats="Н"/></w>
  <w n="4">жалко<ana lemma="жалко" feats="Н"/></w>
 </phr>
 <phr type="ADJ+NOUN"> <w n="5">своих<ana lemma="свой" feats="МС-П: мн,рд"/></w>
  <w n="6">часов<ana lemma="час" feats="С,мр,но: мн,рд"/>
 <ana lemma="часы" feats="С,мн: рд"/></w>
 </phr>
</s>
```

Количество морфологической неоднозначности можно оценить следующим образом. Более 60% словоупотреблений в тексте имеют омонимию граммем (*книги* – им,мн vs. рд,ед), более 30% употреблений имеют омонимию лексем (*хорошо* – Н vs. П,кр,ср; *знакомой* – С vs. П), после локального синтаксического анализа число неоднозначностей сокращается, остается около 10-15% словоупотреблений с лексической омонимией и около 20% с омонимией грамматической, например, в выражении *страницы хорошо знакомой книги* омонимия разбора остается только в слове *страницы*. Дальнейшие методы снятия неоднозначности предполагают использование статистических механизмо, основанных на наличии корпуса "Русский Стандарт" со снятой омнимией.

Литература

1. Андрющенко В. М., 1989. Концепция и архитектура Машинного фонда русского языка. М.: Наука.
2. Вербицкая, Л.А, Казанский, Н.Н., Касевич, В.Б. (2003). Некоторые проблемы создания национального корпуса русского языка. // НТИ, сер. 2, N5.
3. Ершов, А.П. 1981. Методологические предпосылки продуктивного диалога с ЭВМ на естественном языке. // Вопросы философии, №8.
4. Засорина, Л.Н. (ред.), 1977. Частотный словарь русского языка. Наука, Ленинград.
5. Leech, G. 1997. A brief users' guide to the grammatical tagging of the British National Corpus, UCREL, Lancaster University. http://www.hcu.ox.ac.uk/BNC/what/gramtag.html
6. Levontina, I.B., Zalizniak, Anna A. 2001. Human emotions viewed through the Russian language. // Harkins, Jean and Wierzbicka, Anna, (eds.) *Emotions in Crosslinguistic Perspective.*
7. Lönngren, Lennart (ed.), 1993. Частотный словарь современного русского языка. (A Frequency Dictionary of Modern Russian) Acta Universitatis Upsaliensis, Studia Slavica Upsaliensia 32. Uppsala.
8. Sperberg-McQueen, C. M., Burnard, L. (eds.) 2001. Guidelines for Electronic Text Encoding and Interchange. http://www.hcu.ox.ac.uk/TEI/P4X/index.html

Интернет ссылки

1. TEI: Text Encoding Initiative http://www.tei-c.org/
2. БНК: Британский Национальный Корпус http://sara.natcorp.ox.ac.uk/lookup.html
3. БОКР: Большой корпус русского языка, http://bokrcorpora.narod.ru/
4. Диалинг: http://www.aot.ru/
5. МФ РЯ: Машинный Фонд русского языка http://irlras-cfrl.rema.ru/
6. РС: Русский Стандарт http://corpora.yandex.ru/
7. Список омографов Аношкиной http://irlras-cfrl.rema.ru/homoforms/
8. УК: Уппсальский корпус доступен из Университета Тюбингена:
9. http://www.sfb441.uni-tuebingen.de/b1/en/korpora.html
10. Towards a representative corpus of modern Russian