# Представительный корпус русского языка в контексте мирового опыта

С.А. Шаров, РосНИИ Искусственного Интеллекта

*В статье излагаются основные понятия корпусных исследований, предполагающих создание рабочих инструментов (моно- или многоязыковых корпусов текстов) и их использование для исследования лингвистических феноменов, а также приводится обзор наиболее значимых работ в области создания представительных корпусов.*

## 1.Введение

Исследования в области корпусной лингвистики обычно решают сразу две задачи: создание рабочих инструментов и использование этих инструментов для изучения лингвистических феноменов. При этом рабочие инструменты чаще всего предполагают более или менее конкретные лингвистические феномены, которые планируется изучить, а интересы исследователя связаны с возможностями, предоставляемыми рабочими инструментами. Впрочем, подобное взаимопроникновение, можно даже сказать симбиоз доступных технологий и интересов исследователя характерен для многих областей науки.

Создание рабочих инструментов для анализа включает в себя построение моно- и многоязыковых корпусов текстов в рамках области, интересующей исследователя, а также средства разметки корпусов, поиска в них и статистической обработки результатов поиска. Эти ресурсы используются в собственно лингвистических исследованиях, предполагающих анализ языка с экспериментальной точки зрения, т.е. исследование того, какие слова, выражения, грамматические конструкции, типы развития дискурса действительно употребляются носителями языка, как часто и для каких целей.

В настоящее время огромное количество текстов доступно в электронной форме, поэтому часто приходится слышать, что для русского языка собрано множество корпусов, которые используются в лингвистических исследованиях. Эта позиция предполагает, что корпусом является произвольная коллекция текстов по определенной тематике, которые доступны в электронной форме. Назовем такое употребление корпус$_1$. Более рестриктивной является позиция, в соответствии с которой корпус (корпус$_2$) — это коллекция текстов, собранная в соответствии с явно сформулированными принципами и возможно размеченная (annotated) на некотором уровне лингвистического анализа. Это определение соответствует коллекциям текстов, собранным в Машинном Фонде русского языка. Однако в современных исследованиях в области корпусной лингвистики, см. [1], под корпусом (корпус$_3$) часто понимается *представительная* коллекция текстов в смысле корпус$_2$, т.е. корпус$_3$, имеющий конечный размер, может адекватно служить представителем потенциально бесконечного множества текстов некоторого фиксированного типа в некотором

диахроническом срезе. Употребляя слово *корпус* ниже, мы всегда имеем в виду корпус$_3$.

Представительность, в частности, предполагает, что коллекция текстов сбалансирована в отношении жанров и функциональных стилей, и что она имеет достаточные размер и выборку по числу текстов и авторов, чтобы служить основой для статистически достоверных исследований лингвистических феноменов в текстах соответствующей тематики. Типология корпусов соответствует типологии текстов, их составляющих. Корпуса могут быть организованы по жанрам текстов, по виду речи (устная/письменная, диалог/монолог), по авторам и времени создания текстов, по языковому составу текстов, представленных в корпусах, и т.д. В каждом пункте типологии возможна более подробная классификация. Так многоязыковые корпуса могут содержать просто тексты сходной тематики на разных языках (например, корпус путеводителей, написанных на английском и русском языках) или оригиналы и их переводы. В последнем случае тексты могут быть *выровнены* (aligned). Это означает, что множеству, состоящему из некоторых единиц оригинального текста, сопоставлено множество единиц, являющихся в данном контексте переводом, например, разумное выравнивание возможно по абзацам, предложениям или словам. При этом предполагается, что часто соблюдается лишь отношение много ко многим между фрагментами оригинального текста и переводом.

Полноценный обзор различных видов корпусов (включая исторические или многоязыковые корпуса, а также корпуса устной речи) выходит за рамки данной статьи. Отдельной проблемой, которая не затронута в данном описании, является создание корпусов для языков, чья письменность отлична от латинской. Например, для китайского и японского языков проблемой является не только кодировка электронных текстов, но и выделение слов, поэтому в этих языках в проблему превращается даже подсчет числа слов в корпусе, не говоря уже о морфосинтаксической разметке. В данной статье мы хотим остановиться на истории создания представительных корпусов, предназначенных для отражения свойств соответствующего языка в разнообразных жанрах.

## 1.1История создания представительных корпусов

В связи с тем, что компьютер представляет собой удобное средство хранения и обработки больших объемов информации, первые проекты по созданию коллекций текстов и корпусов появились сразу же, как только лингвисты получили доступ к компьютерам с достаточными возможностями для хранения текстов. Первый представительный корпус, Брауновский Корпус (БК, Brown Corpus), был создан в США в 1960-е годы, хотя американские корпусные лингвисты были лучше известны в Европе, чем внутри США [2, 3]. БК состоял из 500 фрагментов объемом по 2000 слов, взятых из текстов, написанных и опубликованных в США в 1961 году [4]. При построении этого корпуса особое внимание обращалось на сбалансированное покрытие различных жанров текстов, для чего была создана иерархия жанров, включавшая, например, элемент СМИ/Репортаж/Спорт, для которого было взято 5 фрагментов из ежедневной прессы и 2 из еженедельной, в то время как для художественной литературы в жанре детективов были взяты фрагменты 4 рассказов и 20 романов (novels). Первая версия корпуса была представлена простым текстовым форматом (с небольшим количеством структурной

разметки для выделения абзацев, заголовков, цитируемых фрагментов и т.п.). Позднее корпус был дополнен разметкой частей речи и морфологических признаков слов.

Результаты экспериментов с этим корпусом привели к заданию стандарта в 1 миллион слов для создания представительных корпусов в разных странах. В частности, по модели близкой к БК в 1970-е годы был создан частотный словарь русского языка [5], построенный на основе корпуса текстов объемом также в 1 миллион слов и включавший 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 разметки), как корпус он недоступен.

Самый известный представительный корпус русского языка также объемом примерно в 1 миллион слов был создан в Университете Упсалы (Швеция), ср. также частотный словарь [6]. Упсальский корпус (УК) состоит из 600 отрывков художественной литературы и информативных текстов, примерно в равной пропорции. Художественная литература взята за период 1960-1988 (отрывки из текстов 40 писателей,), 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 разметка.

Очевидно, что корпус объемом в один миллион слов является недостаточным для адекватного отражения лексических и грамматических свойств языка. Например, выборка из 7 фрагментов общим объемом 3,500 слов считается в БК представительной для спортивных новостей, а выборка объемом 12,000 слов для детективов. По сравнению с детективами выборка научной фантастики в БК состоит из 6 фрагментов объемом в 3000 слов. В то же время слова и грамматические конструкции средней частоты встречаются по несколько раз на миллион слов (со статистической точки зрения язык является большим набором редких событий). Так по современным данным в английском языке частота употребления слов *polite* (вежливый) или *sunshine* (солнечный свет) составляет около 12 раз на миллион слов, что позволяет включить их в число 5000 наиболее употребительных английских слов. Это означает, что при прочих равных условиях они встретятся примерно 12 раз в БК целиком, 0,14 раза в детективах и около 0,04 раза в спортивных новостях и фантастике. Понятно, что корпус такого размера не дает возможности рассматривать употребление таких слов как *polite* или *sunshine* в разных жанрах. Кроме того, статистические методы, рассмотренные ниже, дают надежные результаты, если число найденных результатов составляет по меньшей мере сотни случаев. По этим причинам, а также в связи с ростом компьютерных мощностей, способных работать с большими объемами текстов, в 80-е годы в мире было предпринято несколько попыток создать корпуса большего размера.

В Великобритании такими проектами были Банк Английского (Bank of English) и Британский национальный корпус. В СССР таким проектом был Машинный Фонд русского языка, создававшийся под руководством А.П. Ершова и В.М. Андрющенко [7, 87]. К декларируемым целям создания фонда относились разработка представительного корпуса и подкорпусов различных жанров и соответствующих программных средств, в частности для поддержки разработки компьютерных программ, обрабатывающих естественный язык, а также для комплексной информатизации лингвистических

исследований, включая создание грамматик и словарей. К сожалению, этот проект не завершился созданием собственно представительного корпуса, хотя были собраны коллекции текстов разного типа. В настоящее время из Машинного Фонда доступны некоторые тексты XIX века и газетный корпус за 1997 год.

Похожий проект, начатый в Великобритании в конце 80-х, завершился успешнее: был создан Британский Национальный Корпус (БНК, British National Corpus), который задал новый стандарт создания представительных корпусов, он характеризуется объемом корпуса в 100 млн. слов, использованием полных текстов, подкорпусом устной речи (10 млн. слов), наличием частеречной разметки и доступом через Интернет. Для БНК также была использована подробная классификация документов по нескольким параметрам: виду речи (письменная, устная частная и устная публичная), для письменной по тематике, типам изданий (книги, периодика, машинописные тексты и т.п.), параметрам образования ожидаемой аудитории (высокий, средний или произвольный) и сложности языка (сложный, средний, простой).

По стандарту, заданному БНК, были созданы представительные корпуса многих европейских языков. Более того, характеристика "национальный" в БНК, призванная выделить вариант языка, описываемого корпусом, стала применяться для обозначения представительного корпуса любого языка. По этой модели были созданы, в частности, национальные корпуса испанского, итальянского, хорватского, чешского, а также американского английского, хотя подобный корпус для немецкого и французского до сих пор отсутствует. Для немецкого языка существует COSMAS, коллекция текстов в смысле корпус$_2$, поддерживаемая Институтом немецкого языка.

Второй похожий британский проект Bank of English начал создаваться в 1980-е годы, в 1989 его объем достиг 20 млн. слов, в 2000 - 600 млн. слов. Этот корпус ориентирован на отслеживание изменений в словоупотреблении (monitor corpus) путем регулярного пополнения новыми текстами и сравнения частотных параметров, например, таких как изменение частоты слов и грамматических конструкций, появление новых слов и т.п. Этот корпус служил основой создания словаря Collins COBUILD English Dictionary [9] и ряда английских грамматик, использующих корпусный данные. Еще один вид корпусов представлен корпусом ICE (International Corpus of English), который позволяет сравнивать словоупотребление в различных диалектах английского языка, не только в британском и американском, но и, например, кенийском, новозеландском или сингапурском.

Список представительных корпусов разных языков, доступных через Интернет, приведен в приложении.

## 1.2Разметка корпуса

По мере роста возможностей компьютеров и увеличения количества документов, доступных в электронной форме, размеры корпусов непрерывно росли: от 500 тыс. слов в 1960-е годы до последних проектов, нацеленных на создание корпусов размеров в 1 млрд. слов [10]. Размер является важным параметром оценки корпуса, но не менее важным является состав дополнительной информации внесенной в результате обработки исходного текста.

В связи с тем, что современные средства обработки языка пока не позволяют вносить семантическую или синтаксическую информацию автоматически, а ручная разметка

корпуса объемом в десятки миллионов слов нереальна, разметка большинства больших (и даже не очень больших) корпусов ограничена лишь так называемой лемматизацией (обозначением леммы) и, следовательно, возможностью для пользователя найти в тексте все словоформы произвольной лексемы. Так, например, устроен газетный корпус Лаборатории общей и компьютерной лексикологии и лексикографии (рук. А.А. Поликарпов, см. http://www.philol.msu.ru/~lex/main.htm), корпус прессы и «массовой» литературы А.Н. Баранова и Д.В. Добровольского (см. [11]), русский корпус в университете г. Тампере в Финляндии (авторы Х. Томмола и М. Михайлов, см. [12]. Собственно морфологической разметки в этих корпусах нет, и автоматически искать примеры употребления отдельных грамматических значений в них, соответственно, нельзя.

Более детальная разметка, учитывающая морфологические или морфолого-синтаксические признаки очень трудоемка, потому что лингвистически корректная разметка русских текстов даже на морфологическом уровне без внесения синтаксической информации обязательно требует ручной работы (см. ниже описание эталонного размеченного корпуса). Однако для небольшого корпуса ручная или полуавтоматическая разметка возможна не только для морфологической, но и для синтаксической или семантической информации. Ср., например описание корпуса Санкт-Петербургского университета в статье А.В. Венцова, В.Б. Касевича, Е.В. Ягуновой в настоящем сборнике, ср. также [13]; можно указать также синтаксически размеченный корпус ИППИ (см. [14], а также статью И.С. Чардина в настоящем сборнике), Пражский корпус ([15] или проект Хельсинкского аннотированного корпуса ХАНКО (см. статью А. Мустайоки и М. Копотева в настоящем сборнике); ручная разметка используется и в корпусе FrameNet [16] – по ролям участников ситуации. Одна из первых коллекций синтаксически размеченных текстов Penn Treebank основана на подмножестве текстов Брауновского Корпуса.

В то же время, развитие технологий обработки текстов позволяет вносить все большее количество информации автоматически, в частности, идентифицировать объекты по именам [17] или различать омонимы на основе локального контекста [18, 19]. В первом случае корпус может быть размечен информацией об участниках событий, описываемых в нем, например, именами людей, названиями должностей или фирм, даже если они обозначены различным образом, например, *Mr. Blair, Tony Blair* или *the prime minister*. Во втором случае корпус может быть размечен информацией о конкретном способе использования многозначных слов, например, *table*.

Разметка текста лингвистической информацией в настоящее время чаще всего основана на языке SGML/XML, который предполагает выделение фрагментов текста, например, слово, именная группа, предложение и т.п., и задание значений атрибутов этих фрагментов, например на уровне синтаксических структур с помощью составляющих, ограниченных тэгами <cl> и <phr>:

```
<cl type='finite declarative' function='independent'>
  <phr type='NP' function='subject'>Nineteen fifty-four,
    <cl type='finite relative declarative' function='appositive'>when
      <phr type='NP' function='subject'>I</phr>
      <phr type='VP' function='predicate'>was eighteen years old</phr>
  </cl>, </phr>...
```

Однако SGML/XML задает только синтаксис задания фрагментов и атрибутов, но не конкретный набор, который используется при разметке корпуса. На XML основе в последнее время разработано несколько рекомендаций, наиболее значимыми из

которых являются: EAGLES (European Advisory Group on Language Engineering Standards), TEI (Text Encoding for Interchange), и XCES (XML Corpus Encoding Standard). Правила EAGLES [20] задают общие принципы создания и документирования корпусов и их морфосинтаксической разметки, а также ряд конкретных решений для разметки определенных случаев. В частности, они рекомендуют проводить лемматизацию, но, поскольку лемматизированные корпуса редки, тэгов для лемматизации в EAGLES не предусмотрено. EAGLES также предлагает две возможности для хранения морфологической разметки: каждый признак представлен отдельным атрибутом (POS='NN' number='sing'), или используется сложная морфологическая разметка, в которой цифры соответствуют признакам, например, feats="V3011141101200" означает глагол, 3rd person, singular, finite, indicative, past tense, active, main verb, non-phrasal, non-reflexive (список рекомендуемых признаков и их значений является частью рекомендаций EAGLES). Однако правила EAGLES не содержат готового набора элементов для создания корпуса.

Наиболее разработанным стандартом для собственно лингвистической разметки текстов является XCES [21], который также планируется превратить в международный стандарт (ISO TC37/SC4). XCES задает абстрактную *мета*модель, которая обеспечивает средства создания всех разумных моделей лингвистических разметок, удовлетворяющих правилам EAGLES. Для этого определены абстрактные тэги узлов <struct> и их признаков <feat>. Для каждого узла должен быть задан его тип, например, p-level, s-level, w-level, m-level, для абзацев, предложений, слов и морфем. Это позволяет представлять мультислова, как одну единицу анализа, например, *as well as* в английском или глаголы с отделяемыми приставками, например, *zunehmen* в немецком. Можно представлять и варианты неоднозначного разбора, включая более и менее вероятные варианты. Можно также проводить декомпозицию одного слова в пределах разметки, например, для *zum* как *zu dem* в немецком, но это достигается за счет вложенной системы кодов:

```
<struct type=”W-level”>
  <seg target=”#w1”/>
  <struct type=”W-level”>
        <feat type=”lemma”>zu</feat>
        <feat type=”pos”>PREP</feat>
  </struct>
  <struct type=”W-level”>
        <feat type=”lemma”>dem</feat>
        <feat type=”pos”>ART</feat>
        <feat type=”case”>dat</feat>
        <feat type=”number”>sing</feat>
  </struct>
</struct>
```

Большинство корпусов в настоящее время не используют чрезмерно сложного механизма XCES, а используют набор тэгов TEI [22]. Хотя стандарт TEI не ориентирован на лингвистические задачи и хранение размеченных корпусов, он определяет и основные способы лингвистической разметки, которые используются во многих корпусных проектах, например, использование тэгов <w> для обозначения слов, <s> для предложений, <phr> для групп и т.д.

## 1.3Использование корпусов для лингвистических исследований

Любой корпус создается как средство отражения и эмпирического исследования явлений, встречающихся в языке (или подъязыке, в случае специализированного

корпуса). Наличие компьютерного корпуса не меняет радикально деятельность лингвиста. Корпуса текстов использовались для создания грамматик и словарей задолго до появления компьютеров. Словарь английского языка Сэмюеля Джонсона (середина 18-го века) и грамматика Отто Есперсена (первая половина 20-го века), а равно и академические грамматики и словари русского языка создавались на основе анализа реальных примеров словоупотребления, записанных на карточках. Однако, компьютерный корпус служит инструментом, с помощью которого можно проводить исследования, методологически отличные от традиционных.

Создание представительных корпусов и их разметка на различных уровнях влечет за собой создание словарей и грамматик, построенных на основе корпусной методологии. Представительный британский корпус LOB, построенный по модели Брауновского корпуса, лег в основу [23] Quirk, et al, 1985). Создание первого большого английского корпуса Bank of English (20 млн. словоупотреблений в 1980-е годы) привело к созданию серии грамматик и словарей серии COBUILD [24, 25, 26]. Создание БНК привело к появлению новой грамматики [27]. Задача, встающая перед корпусно-ориентированной грамматикой, заключается в создании описания, которое может дает адекватный анализ для *любого* словоупотребления, зафиксированного в корпусе, а наибольшее внимание обращается на наиболее частотные случаи.

Непосредственные наблюдения, которые можно провести над текстом корпуса, касаются лексических характеристик, таких как частота встречаемости отдельных слов и словосочетаний, см. [28]. При поиске устойчивых словосочетаний наиболее просто (и лингвистически интересно) подсчитать Т-статистику по Стьюденту и взаимную информацию, а также их произведение. Критерий взаимной информации указывает то, насколько совместное употребление двух слов далеко от случайного:

$$(1) \qquad \mathrm{MI_{ij}} = \log_2\left(\frac{O_{ij}}{E_{ij}}\right); \quad O_{ij}=\frac{F_{ij}}{N}\ ;\ E_{ij}=\frac{F_i}{N}\times\frac{F_j}{N}$$

где $F_i$ число употреблений слова в корпусе, $F_{ij}$ число случаев, где они употреблены вместе, $N$ число слов в корпусе, $O_{ij}$ относительная частота случаев, когда слова встречаются вместе (observed frequency), $E_{ij}$ ожидаемая частота случайного распределения (expected frequency). Эта статистика находит слова, которые относительно редко встречаются раздельно, даже если это единичные случаи. Например, *бельгийский кольт* и *тюремная роба* являются одними из самых значимых словосочетаний типа СУЩ+ПРИЛ в корпусе лингвистических статей (Труды семинара Диалог, объем примерно 700,000 слов), поскольку эти слова крайне редки в этом корпусе (каждое один раз встречается по отдельности и один раз вместе в составе словосочетаний). Т-статистика, наоборот, находит наиболее частые словосочетания:

$$(2) \qquad \mathrm{T_{ij}} = \frac{O_{ij}-E_{ij}}{\sqrt{O_{ij}}}\ ;$$

При этом Т-статистика отличается от простого упорядочивания словосочетаний по частоте. Например, в корпусе Диалога словосочетание *описание языка* встречается чаще, чем *часть речи* (244 против 136), но второе словосочетание имеет больший вес по Т-статистике, поскольку и *описание,* и *язык* слишком частотны в этом корпусе (считается разность между ожидаемой и реальной частотой встречаемости). С другой стороны, в отличие от статистики взаимной информации Т-статистика находит частые

терминологически незначимые словосочетания: *большое количество, данная работа, следующим образом* и т.п. В качестве более точной меры описания близости слов можно предусмотреть либо более сложные статистические методики, либо простое перемножение критериев Т-статистики и взаимной информации, которое позволяет исключить hapax legomena и частые сочетания. Введение в набор статистических средств, используемых при обработке корпусов, см. в [17].

## 2.К созданию представительного корпуса русского языка

Как уже было сказано, состояние дел в российской корпусной лингвистике можно кратко охарактеризовать отсутствием представительного корпуса современного русского языка. В настоящее время идея осуществления такого проекта возрождается (см. статью Л.А. Вербицкой, Н.Н. Казанского и В.Б. Касевича в настоящем сборнике). Предполагаемый корпус будет иметь разные составляющие. Отдельной задачей видится, например, представительный корпус XIX-го и первой половины XX-го века, по понятным причинам состоящий из письменных, прежде всего, художественных текстов; специфических проблем создания такого корпуса (и тем более трудностей, связанных с созданием корпуса древнерусских текстов) мы в настоящей статье касаться не будем.

Высказываемые ниже соображения применимы прежде всего к корпусу современного русского языка (со второй половины XX-го века). Они основаны на собственном экспериментальном проекте автора, который мог бы быть использован для дальнейшей коллективной работы в этой области. Этот проект фигурирует пока под условным названием БОКР (Большой корпус русского языка). БОКР представляет все значимые виды использования русского языка в настоящее время. Автор проекта предполагал, что относительно небольшой фрагмент этого корпуса, наилучшим образом отражающий нейтральный литературный язык второй половины XX в., мог бы быть полностью размечен морфологически – со снятой вручную омонимией (тогда как разметка остальной части корпуса будет произведена автоматически). Об этом фрагменте ниже мы будем говорить как об эталонном размеченном корпусе русского языка. Отметим, что задел для создания эталонного корпуса тоже существует; подробнее о проблемах, связанных с его созданием см. [29]. Более подробная сравнительная характеристика БОКР и его размеченного фрагмента приведена в Таблице 1.

### 2.1Типология текстов

При создании представительных корпусов требуется максимально широкое покрытие различных типов текстов и функциональных стилей, для чего была разработана типология текстов на основе рекомендаций Синклера [24]. Англоязычная литература содержит множество предложений по классификации для создания представительных корпусов, например, [30, 31, 32, 33] и т.д. Принимая эти рекомендации необходимо иметь в виду, что среда использования языка оказывает существенное влияние на классификацию текстов. Так очевидное несоответствие между российской и англоязычной культурами выражено и на уровне жанров. Такие жанры как выделяемые в БК Foundation reports или Popular Lore с трудом допускают даже адекватный перевод на русский.

| | Размеченный эталонный корпус | БОКР |
|---|---|---|
| размер | 10 млн. слов (1000 текстов) | 100 млн. слов (20000 текстов) |
| качество | сбалансированная подборка современной художественной и мемуарной прозы | сбалансированный набор текстов, пред ставляющих все значимые виды исполь зования языка в настоящее время |
| разметка | морфологическая (с элементами синтаксической) с ручным снятием омонимии | морфологическая (с элементами синтаксической) с автоматическим снятием омонимии |
| доступ | свободный доступ через Интернет с единым языком запросов | |

Таблица 1. Параметры корпусов

С другой стороны, классификация, предлагаемая Синклером, основана на логических свойствах коммуникации и может быть адаптирована для описания русского дискурса. Синклер выделяет два класса факторов, влияющих на выбор текстов в корпусе: внешние (E), внеязыковые факторы, которые могут повлиять на структуру или содержание текста, и внутренние (I), факторы, отражающие свойства языка, используемого в тексте. Выделяются три группы E-факторов:

1. E1 (origin) - факторы, относящиеся к созданию текста автором;
2. E2 (state) - факторы, относящиеся к внешним признакам текста;
3. E3 (aims) - факторы, относящиеся к целям создания текста и его влиянию на аудиторию.

Два основных I-фактора:

1. I1 (topic) - предметная область текста;
2. I2 (style) - стилистические особенности (частично зависящие от E-факторов).

К группе **E1** (параметры создания текста) относятся, в первую очередь, время создания текста и возраст автора на этот момент, пол автора и регион происхождения автора. Для региона важна грубая классификация на столичный (Москва и Санкт-Петербург), европейский, сибирский и южный, для возраста на детский, молодежный, взрослый и пожилой.

Стремление отразить современный русский язык ограничивает диахронические параметры 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 [5] 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 историческая 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 с 1990 (это ограничение объяснимо и техническими причинами: более ранние тексты малодоступны в электронном виде), а газеты и журналы берутся с 1995.

Для описания текста по его внешним признакам (**E2**) предлагается иерархия, отличающаяся от традиционной, в первую очередь, наличием четырех режимов речи: устной, письменной, письменной, предназначенной для произнесения вслух, и электронной коммуникации. Последняя подобна устной речи спонтанностью порождения (подобно телефонному звонку или очной дискуссии), но она все равно остается письменной (в частности, отсутствует просодическая информация).

Среди параметров письменной речи выделяются печатные издания, подразделяемые на книги, периодику и брошюры, а также переписка разного рода и машинописные тексты. Устную речь можно подразделять на записанную в естественных условиях, в студии и телефонные разговоры.

Группа факторов **E3** касается целей создания текста и его влияния на аудиторию. К параметрам аудитории, которые оказывают существенное влияние на текст, отнесены ее размер, близость аудитории говорящему и ограничения на пол, возраст и уровень образования аудитории. По размеру аудитории речь делится на публичную (более 50 читателей/слушателей, с подклассами сотни, десятки тысяч и миллионы) и частную, в свою очередь подразделяемую на личную (2 участника), небольшую группу (до 5), группу средних размеров (до 20) и коллектив. По параметру близости в большинстве случаев публичная аудитория деперсонализирована. Если говорящий/пишущий может описать каждого участника коммуникации, их близость классифицируется по шкале: хорошее личное знакомство, личное знакомство и его отсутствие.

Под целями создания текста понимается коммуникативная функция текста:

обсуждение (аргументация, полемика, изложение позиции и т.п.)

рекомендации (отчеты, предложения, законы и т.д.)

развлечение

Сюда входят различные жанры художественной литературы, а также биографические и автобиографические тексты, дневники и мемуары.

учебные (в эту категорию входят как школьные или вузовские учебники, так и практические рекомендации).

информативные (в эту категорию входят только те тексты, целью которых является исключительно предоставление информации и которые не могут быть включены в другие категории, например, энциклопедии и справочные пособия)

При построении корпуса не слишком важна глубина кодирования предметной области, затрагиваемой текстом (фактор **I1**), поскольку корпус не является универсальной энциклопедией. Кроме того, общие классификации, подобные УДК, редко применимы к тексту и в еще меньшей степени применимы к устной речи, поскольку сколько-нибудь значимый отрезок текста может затрагивать несколько предметных областей одновременно. При построении корпуса можно иметь грубую классификацию, выделяющую естественные и общественные науки, политику и экономику, искусство и досуг, см. обсуждение предметных областей в БНК и БОКР в Таблице 2.

| Жанр | Объем в БНК | Объем в БОКР |
|---|---|---|
| Устная речь | 10,7 % | 5 % |
| Художественная литература | 16,7 % | 30 % |
| Политика | 18,9 % | 15 % |
| Экономика | 7,6 % | 5 % |
| Естественные науки | 3,8 % | 5 % |
| Гуманитарные науки | 14,2 % | 12 % |
| Религия и философия | 3,1 % | 3 % |
| Искусство | 6,8 % | 5 % |
| Досуг | 11,2 % | 10 % |
| Прикладные области | 7,2 % | 10 % |

Таблица 2. Сравнительный анализ жанров в БНК и БОКР

Система кодирования стилистических особенностей текстов (факторы **I2**) пока не разработана. Наиболее очевидным кажется выделение таких отклонений от нейтрального стиля как формальный, академический и просторечный. Для художественной литературы представляется полезным деление текстов на представляющие стандартный литературный язык (например, Ю. Трифонов), сниженный язык (Ю. Алешковский, Э. Лимонов), язык с имитацией региональных особенностей («деревенская проза»), выраженно индивидуальный авторский язык, отличный от нормы (Саша Соколов) и др. Конечно, приведенные примеры авторов носят лишь иллюстративный характер, поскольку описывается основной стиль каждого текста.

## 2.2Принципы построения представительных корпусов

Вышеописанные пять факторов классификации составляют основу для балансировки коллекции текстов, включаемых в БОКР. Мы стремимся представить в корпусе каждое значение этих факторов, которые релевантны для задач, решаемых этим корпусом.

Важным техническим параметром является также размер текста. Корпус большого размера проще собрать из длинных текстов, но поскольку длинных текстов в таком корпусе оказывается относительно мало, идиосинкразическое использование языка в каждом из них оказывает существенное влияние на характеристики языка, описанные корпусом. По этой причине в двух корпусах предполагается соблюсти баланс между размером и количеством текстов в пользу более коротких текстов.

Поскольку в БОКР должны быть представлены все функциональные жанры русской речи, должны быть учтены все возможные комбинации параметров, хотя количество текстов в каждой группе зависит от количества соответствующих текстов в русском дискурсе и их доступности в электронном виде.

Возьмем в качестве примера набор текстов, ограниченный предметной областью "Политика", подобласть "Внутренняя политика" (параметр I1). Вариация по другим параметрам включает в себя: тексты, написанные в нейтральном, формальном,

академическом и просторечном стиле (I2), созданные мужчинами, женщинами или авторскими коллективами за период 1995-2000 годов в различных регионах России (E1), опубликованные в газетах, журналах, книгах и машинописных отчетах (E2, для устной речи можно выделить публичное обсуждение политических событий в теле- и радиоэфире и в частных беседах), предназначенные для аудитории разного размера, социального положения и уровня знаний о теме текста (от отчета, написанного для президента, до статьи в Московском Комсомольце), а также предназначенные для выполнения разных коммуникативных целей (обсуждения, рекомендации, обучения или развлечения).

Любой текст, предназначенный для включения в корпус, должен быть описан в рамках этих базовых параметров. В качестве примера текста в области внутренней политики возьмем текст Конституции РФ: это текст, написанный в формальном стиле (I2) в 1993 в Москве, личное авторство отсутствует (corporate) (E1), это письменный текст, опубликованный в виде книги объемом 9500 (E2), предназначенный для очень большой аудитории без ограничений на образование (даже если он не был прочитан подавляющим большинством населения), цель создания: рекомендация, подтип: юридический документ.

Некоторые комбинации параметров являются крайне маловероятными: например, книги, написанные мужчинами в формальном стиле в области естественных наук для массовой женской аудитории с целью развлечения (комбинация формального стиля изложения и развлекательных целей, а также конкретного пола предполагаемой аудитории и тематики естественных наук кажутся маловероятными). Некоторые параметры взаимно исключают друг друга, например, машинописный отчет, предназначенный для миллионной аудитории или личное обсуждение по телевидению. Все же любое *разумное* сочетание параметров должно быть при возможности представлено в корпусе несколькими текстами.

Что касается доступности текстов в электронном виде, некоторые типы текстов легко доступны, например, художественная литература и газетно-журнальные тексты. Другие виды текстов так называемых "эфемерных жанров", например, деловую или личную переписку или устную речь, гораздо сложнее получить и использовать при создании корпусов, хотя эти жанры весьма важны для построения представительного корпуса, поскольку они отражают реальное использование современного русского языка подавляющим большинством его носителей. В связи с этими проблемами, при создании БОКР делается попытка при возможности представить в корпусе максимальное количество текстов "эфемерных жанров", поскольку электронные версии художественной литературы и газетно-журнальных текстов слишком легко превзойдут их по объему. Предварительная оценка пропорций различных жанров может основываться на опыте БНК (при этом около 4,5% письменных текстов в БНК составляют неопубликованные источники, главным образом, эфемерных жанров). Предварительная оценка баланса БОКР в сравнении с БНК приведена в Таблице 2. Наиболее существенное отличие касается доли художественной литературы, поскольку она составляет основу для определения нормы русского языка и имеет существенно большее влияние на русский язык по сравнению с английской традицией. Более того, наш эталонный корпус описывает исключительно русский литературный язык.

В отличие от БОКР, который предполагает использование всех видов художественной литературы, включая детектив, фантастику, исторические,

приключенческие, любовные и юмористические тексты, эталонный корпус отражает норму современного русского литературного языка, определенную на основе текстов, написанных в основном после 1960 года. Поэтому в него предполагается включить в первую очередь художественную прозу, соответствующую стандартному русскому литературному языку. Эталонный корпус, как и БОКР, включает также примеры драматургии, биографий, мемуаров и опубликованной переписки профессиональных писателей. Следует особо отметить, что набор текстов, включаемых в эталонный корпус, не определяется непосредственно художественными достоинствами произведений и создание корпуса не преследует литературоведческих целей, хотя может представлять некоторый интерес и для литературоведов. При этом, в связи с ориентацией корпуса на стандартный литературный язык, предполагается отбор текстов по следующему формальному параметру: лексический состав (за исключением имен собственных) не должен существенно отличаться от стандартного языка, зафиксированного в современных словарях и в системе морфологического анализа, использованной для разметки текста. В первую очередь, корпус призван отразить лексико-грамматические характеристики, которые активно использовались писателями с соответствующей вариацией по авторам, стилям, времени создания, размерам текстов и т.п. Вариация по времени создания должна обеспечить наличие в корпусе текстов по всему временному срезу от 1960 до 2000 года; аналогично для размера: должны быть в пропорциональном объеме представлены рассказы, повести и романы.

## 2.3Разметка текстов

Корпус предполагает проведение лемматизации и морфосинтаксической разметки текстов. Хотя многие корпуса английского языка ограничивались частеречной разметкой, в случае русского языка безусловно необходима лемматизация (иначе будет затруднен поиск многих словоформ) и приписывание граммем (это позволит исследовать предложное и глагольное управление). Поскольку выделение именных и предложных групп в русском языке автоматическими методами достаточно надежно, корпус предполагает также частичную синтаксическую разметку.

После этапа морфологического анализа словоформы с неоднозначными грамматическими характеристиками (например, род, число, падеж) составляют около 60% словоупотреблений, словоформы с неоднозначным выделением лексемы и части речи – около 30%. Однако частичный синтаксический анализ на уровне именных и предложных групп может оказать существенную помощь при разрешении некоторых видов неоднозначности, в частности:

- неоднозначности падежных форм, например, *новой книги*, где при неоднозначности отдельно взятых словоформ у целой именной группы возможен только родительный единственного;
- неоднозначности субстантивированных прилагательных: *старший* в *мой старший брат* получает только интерпретацию прилагательного (а *мой* – местоимения), но в *старший группы* – только существительного;
- неоднозначности притяжательных и личных местоимений, например, *в его книге,* где невозможна интерпретация личного местоимения,
- омонимии между существительным или прилагательным и глаголом в повелительном наклонении: *кривей, мой, полей,* при наличии в непосредственной

близости другого, однозначно определяемого, глагола, глагольная форма невозможна.

Наоборот, при наличии согласования с близлежащим существительным или местоимением в именительном падеже формы *были, замер, стали*, однозначно интерпретируются как глаголы. Аналогично можно разрешить и многие случаи омонимии между кратким прилагательным и наречием.

Оставшуюся неоднозначность простыми средствами снять невозможно. Особенно часты случаи омонимии между существительными внутри части речи, например, *поле* является формой *пол, поле* и *пола* (также женского имени *Поля* при написании с большой буквы), которые во многих случаях можно снять только полным синтаксическим анализом на уровне предложения, что сделать автоматически в настоящее время невозможно. Более того, без семантического и прагматического анализа полного текста нельзя решить, что имелось в виду в заголовке газетной статьи "*Не храните свои деньги в банке*". Часты также случаи неоднозначности между наречиями и предикативами, например, *жутко, забавно, хорошо*, или союзами и частицами (*будто, лишь, словно*). При использовании частичного синтаксического анализа остается примерно 15% форм, неоднозначных по лемме и части речи, и 20% форм, неоднозначных по грамматическим признакам.

В размеченном эталонном корпусе оставшиеся виды неоднозначности предполагается снимать вручную (примерная оценка скорости разрешения неоднозначности — 300 тыс. слов за человекомесяц). В БОКР некоторые решения принимаются на основе частоты форм, например, *спина* почти всегда анализируется как существительное женского рода, а *гноме, методе* как мужского рода в предложном падеже (за редких исключением случаев, когда построена именная группа, определившая другой род, например, *о знаменитой гноме*). Неоднозначность между наречиями и предикативами или союзами и частицами представлена в виде совмещенного признака: Н/ПРЕДК, СОЮЗ/ЧАСТ. Часть неоднозначности по выбору леммы и части речи остается (например, для словоформ *банки, часы*) и составляет примерно 3.6% словоупотреблений.

Морфосинтаксическая разметка основана на формате TEI с помощью тэгов <s>, <phr>, <w>. Поля с анализом различного вида присваиваются фрагментам с помощью тэга <ana> (*ana*lysis). Это позволяет представлять одинаковым образом как однозначные, так и омонимичные разборы:

```
<s n="1">
   <w n="1">Они<ana lemma="они" feats="МС,мн,3л: им"/></w>
   <w n="2">шли<ana lemma="идти" feats="Г,нс,нп,дст: мн,прш"/>
                        <ana lemma="слать" feats="Г,нс,пе: ед,дст,2л,пвл"/>
   </w>
    <w n="3">домой<ana lemma="домой" feats="Н"/></w>
</s>
```

## 3.Вместо заключения

Подавляющее большинство исследований в области корпусной лингвистики начиналось на материале английского языка. Причиной этого является не только и не столько активное развитие компьютерной техники в США а интеллектуальный климат в Британской лингвистике в 60-80 годы 20 века. В США в это время властвовал хомскианский подход, основанный на лингвистической интуиции, которая не требует

наличия корпусных данных (зачастую хаотичных и зависящих от более широких контекстов высказывания), поскольку объектом изучения является возможность построения правильных языковых конструкций (well-formedness), а различие между правильными и неправильными конструкциями может быть проведено любым носителем изучаемого языка. В противоположность рационалистскому подходу, основанному на лингвистической интуиции, проводящей различие между правильными и неправильными конструкциями, эмпирический подход предполагает, что язык является ресурсом, обеспечивающим набор возможностей для коммуникации. Этот набор реализуется в дискурсе, поэтому объектом исследования в лингвистике является результат реализации этого ресурса, а именно слова и конструкции, употребленные в тексте.

В отличие от США, в британской лингвистике были сильны эмпирические тенденции, которые предполагали использование реальных примеров для проверки лингвистических гипотез, отметим, в первую очередь, исследования Джона Фёрса и его учеников Грегори, Синклера, Хэллидея и др. Это и привело к созданию многих корпусов и разработке корпусных исследований на материале английского языка. Эмпирический подход явно присутствовал и в нашей лингвистике, в которой лингвистический анализ практически всегда сопровождался примерами реального словоупотребления. Просмотр публикаций по корпусной лингвистике (например, в журнале The International Journal of Corpus Linguistics) показывает, что многие темы, поднимаемые исследователями в этой области, созвучны с классическими темами российской лингвистики, посвященными исследованию того, какие слова, выражения, грамматические конструкции и типы развития дискурса действительно употребляются носителями языка и в каких контекстах.

## Литература


1. McEnery T, Wilson A. *Corpus Linguistics.* Edinburgh: Edinburgh University Press, 2nd edition, 2001.
2. Fillmore, Ch.J. 'Corpus linguistics' or 'Computer-aided armchair linguistics'. In: J. Svartvik (ed.). *Directions in Corpus Linguistics.* Berlin: de Gruyter, 1992, p. 35-60.
3. Teubert, W. Corpus Linguistics and Lexicography. *International Journal of Corpus Linguistics.* Special issue, 2001, p. 125-153.
4. Kučera, H., Francis, W.N. *Computational analysis of present-day American English.* Providence: Brown University Press, 1967.
5. Засорина, Л.Н. (ред.). *Частотный словарь русского языка.* Л.: Наука, 1977.
6. Lönngren, Lennart (ed.). *Частотный словарь современного русского языка.* (A Frequency Dictionary of Modern Russian) Acta Universitatis Upsaliensis, Studia Slavica Upsaliensia 32. Uppsala, 1993.

7. Ершов, А.П. Методологические предпосылки продуктивного диалога с ЭВМ на естественном языке // *Вопросы философии,* 1989, № 8.

8. Андрющенко В. М. *Концепция и архитектура Машинного фонда русского языка.* М.: Наука, 1989.

9. Sinclair J. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press, 1991.

10. Cieri, C., Liberman, M. Language resources creation and distribution at the Linguistic Data Consortium // *Proc. of Language Resources and Evaluation Conference (LREC02)*, 2002, p. 1327-1333.

11. Баранов А.Н., Михайлов М.Н., Сидоров Г.О. "Динамический корпус текстов" как новая технология прикладной лингвистики // Труды международного семинара Диалог '98 по компьютерной лингвистике и ее приложениям. Т.2. 1998.

12. Михайлов М.Н. Контекстно-свободная лемматизация как временное решение насущных проблем // Алфавит. Смоленск, СПГУ, 2002, с. 197-210.

13. Венцов А.В., Касевич В.Б. Словарь для модели восприятия речи // Вестник С.-Петербургского ун-та. 1998. Сер. 2. Вып. 3.

14. Богуславский И.М., Григорьев Н.В., Григорьева С.А., Иомдин Л.Л., Крейдлин Л.Г., Санников В. З., Фрид Н.Е. *Аннотированный корпус русских текстов: концепция, инструменты разметки, типы информации*. // Труды Международного семинара Диалог-2000.

15. Hajičová E., Panevová J., Sgall P. Language Resources Need Annotations To Make Them Really Reusable: The Prague Dependency Treebank // *Proc. of Language Resources and Evaluation Conference*, 1998, p. 713-718.

16. Fillmore, Ch.J., Baker, C.F., Sato, H. The FrameNet Database and Software Tools // *Proc. of Language Resources and Evaluation Conference (LREC02)*, 2002.

17. Manning, C.D., Schuetze, H. *Foundations of Statistical Natural Language Processing.* MIT Press: Cambridge, MA, 1999.

18. Kilgarriff, A., Rosenzweig, J. English Senseval: Report and Results. // *Proc. of Language Resources and Evaluation Conference (LREC00)*, 2000, p. 1239-1244.

19. Kilgarriff, A. Web as Corpus // *Proc. of Corpus Linguistics Conference*. April, 2001, Lancaster, UK, 2001.

20. EAGLES: *Recommendations for the morphosyntactic annotation of corpora*, EAG-TCWG-MAC/R. 1996. Available from ftp://ftp.ilc.pi.cnr.it/pub/eagles/corpora/annotate.ps.gz

21. Ide, N., Romary, L. Standards for language resources. In *Proc. of Language Resources and Evaluation Conference (LREC02),* May 2002. Las Palmas, Spain, 2002, p. 59-65.

22. Sperberg-McQueen, C. M., Burnard, L. (eds.). *Guidelines for Electronic Text Encoding and Interchange*, 2001. http://www.hcu.ox.ac.uk/TEI/P4X/index.html

23. Quirk, R., Greenbaum, S., Leech, G., Svartvik J. *A Comprehensive Grammar of the English Language*. London: Longman, 1985.

24. CCEG: *Collins COBUILD English Grammar.* The COBUILD Series from The Bank of English. London: Harper Collins Publishers, 1990.

25. Sinclair J. *Preliminary recommendations on text typology.* EAGLES Document EAG-TCWG-TTYP/P, 1996. http://www.ilc.pi.cnr.it/EAGLES96/texttyp/texttyp.html

26. Hunston, S., Francis, G. *Pattern grammar: a corpus driven approach to the lexical grammar of English*. Amsterdam: John Benjamins, 1999.

Lee, D. Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology* Vol. 5, No. 3, September 2001, pp. 37-72. http://llt.msu.edu/vol5num3/pdf/lee.pdf

27. Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. *The Longman Grammar of Spoken and Written English.* London: Pearson Education, 1999.

28. Шайкевич, А.Я. Дифференциальные частотные словари и изучение языка Достоевского (на примере романа "Идиот") // *Слово Достоевского*, Москва: ИРЯ РАН, 1996, с. 195-253. Также http://irlras-cfrl.rema.ru:8101/publications/ryano.htm

29. Сичинава, Д.В. К задаче создания корпусов русского языка // *НТИ, сер. 2,* 2002, N 12.

30. Biber, D. *Variation across speech and writing.* Cambridge, UK: Cambridge University Press, 1988.

31. Biber, D. *Dimensions of register variation: A cross-linguistic comparison.* Cambridge, UK: Cambridge University Press, 1995.

32. Halliday, M.A.K.; Hasan, R. *Language context and text: Aspects of language in a social-semiotic perspective.* Oxford, UK: Oxford University Press, 1985.

33. Hoey, M. *On the surface of discourse.* London: Allen and Unwin, 1983.

**Интернет ссылки на корпуса**

В предлагаемом списке представлены представительные корпуса разных языков, доступные через Интернет. Представительная и обновляемая коллекция Интернет ресурсов представлена в направлении "Корпусная лингвистика" на сайте http://www.dialog-21.ru

**Английский язык**

БА: Банк Английского, http://titania.cobuild.collins.co.uk/

БК: Брауновский Корпус, http://www.hd.uib.no/icame/brown/bcm.html

БНК: Британский Национальный Корпус, http://sara.natcorp.ox.ac.uk/lookup.html

ICE: International Corpus of English, http://www.ucl.ac.uk/english-usage/ice/

Penn Treebank: http://www.cis.upenn.edu/~treebank/

**Русский язык**

БОКР: Большой Корпус русского языка, http://bokrcorpora.narod.ru/

Эталонный размеченный корпус: http://corpora.yandex.ru/

УК: Уппсальский корпус, доступен из Университета Тюбингена:

http://www.sfb441.uni-tuebingen.de/b1/en/korpora.html

Полезные ссылки для русского:

МФ РЯ: Машинный Фонд русского языка, http://irlras-cfrl.rema.ru/

Лаборатория общей и компьютерной лексикологии и лексикографии (рук. А.А. Поликарпов) http://www.philol.msu.ru/~lex/main.htm

Диалинг: морфологический анализатор, http://www.aot.ru/

Частотный список современного русского языка: http://bokrcorpora.narod.ru/frqlist/frqlist.html

**Другие языки**

Болгарский: http://www.lml.bas.bg/

Итальянский: http://www.cilta.unibo.it

Китайский: http://www.sinica.edu.tw/ftms-bin/ftmsw3

Немецкий: http://corpora.ids-mannheim.de/~cosmas/

NEGRA - синтаксически аннотированный корпус немецкого языка: http://www.coli.uni-sb.de/sfb378/negra-corpus/

Португальский: http://acdc.linguateca.pt/

Хорватский: http://www.hnk.ffzg.hr/

Чешский: http://ucnk.ff.cuni.cz/
The Prague Dependency Treebank - синтаксически аннотированный корпус чешского языка: http://ufal.mff.cuni.cz/pdt/pdt.html

Эстонский: http://psych.ut.ee/gling/en/corpusb/

**Ссылки по истории создания корпусов:**

http://www.uni-koeln.de/phil-fak/englisch/bald/corpora.htm

http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/history.html