А.М. Лаврентьев
Институт филологии СО РАН, Новосибирск

Корпусная лингвистика: идеология, методы, технологии

Термин *корпусная лингвистика* сейчас весьма популярен. Составление корпуса русских текстов включено в число приоритетных направлений работы РАН. Создание корпусов текстов рассматривается рядом ученых как важнейшая гуманитарная задача лингвистики[1]. Вместе с тем место корпусной лингвистики среди направлений языкознания и само понятие корпуса не всегда понимаются одинаково. Методика корпусных исследований, существующие ресурсы и средства работы во многом по-прежнему остаются terra incognita. В данном обзоре мы постараемся привести определение наиболее важных терминов, использующихся в корпусных исследованиях, осветить актуальные методологические и технологические проблемы этого направления и представить несколько доступных исследователям корпусов и коллекций текстов.

1. Что такое корпусная лингвистика?

К сфере корпусной лингвистики можно отнести все лингвистические исследования, опирающиеся на материал корпуса текстов. Определение корпуса мы постараемся дать несколько позже, а пока отметим, что корпусная лингвистика – это не направление, связанное с определенным ярусом языковой системы (как фонетика, лексикология или синтаксис), или определенной теорией (как функциональная или генеративная грамматика), или аспектом анализа (формальным, семантическим или прагматическим). Это скорее идеология, согласно которой результаты лингвистического исследования должны опираться прежде всего на анализ текстов (устных или письменных), а не на интуицию исследователя или информанта.

По-видимому, найдется не много сторонников радикального подхода, полностью отрицающих роль интуиции. Для лингвистов, причисляющих себя к корпусному направлению, речь идет именно о системе приоритетов: любой вывод должен подтверждаться материалом «естественных» текстов, а не только суждениями о приемлемости той или иной конструкции, полученными в условиях лингвистического эксперимента.

2. Что такое корпус?

В известном смысле подавляющее большинство современных лингвистических исследований (за исключением чисто абстрактных теорий типа глоссематики или раннего генеративизма) так или иначе опирается на текстовый материал. Наверное, всем лингвистам приходилось работать с карточками или с электронными записями (транскрипциями) текстов. Если выводы исследования целиком опира-

1 Эта мысль была высказана, в частности, В.А. Плунгяном в докладе на международной конференции «Формирование образовательных программ, направленных на создание нового типа гуманитарного образования в условиях полиэтнического сибирского сообщества» (Новосибирск, 28-30 октября 2003 г.).

ются на четко определенный текстовый материал, этот материал можно назвать корпусом. Вопрос лишь в том, насколько данный корпус показателен (репрезентативен) для того, чтобы судить о языке в целом.

Принято проводить различие между корпусом и коллекцией (или библиотекой) текстов. В качестве характерных черт корпуса нередко называют большие размеры (десятки миллионов словоупотреблений) и наличие лингвистической разметки.

На наш взгляд, отличительным признаком корпуса является, прежде всего, **репрезентативность**. При этом размеры корпуса, отвечающего требованию репрезентативности, зависят от того, для каких исследований он предназначен. Для исследований в области фонетики, просодики, морфологической типологии («индексы Гринберга»), определения доминирующего порядка слов и наиболее частотных синтаксических моделей и т.п. нет необходимости в привлечении огромных массивов текстов. Репрезентативность здесь будет определяться представленностью различных функциональных стилей, диалектов и социолектов, диахронической перспективой. Впрочем, корпус конкретного исследования вполне может быть ограничен рамками одного регионального или социального диалекта или даже речевой продукцией отдельной личности.

В то же время, если нас интересуют какие-то периферийные явления лексики или грамматики, процессы грамматикализации отдельных лексем, возникновение и развитие определенных синтаксических конструкций, необходимо привлечение значительно более широкого материала.

3. Reference corpus

В идеале нужно стремиться к созданию Корпуса Языка (с большой буквы) – корпуса, «репрезентативного во всех отношениях», который мог бы служить надежным источником данных для любых лингвистических исследований.

В англоязычной литературе такой корпус обозначается термином *reference corpus* ‘образцовый (?) корпус’. Английский ученый Дж. Синклер, автор программной статьи о типологии корпусов, дает следующее определение: «Образцовый корпус создается для того, чтобы предоставить полную информацию о языке. Он должен быть достаточно крупным, для того чтобы репрезентировать все существенные разновидности этого языка и его характерных пластов лексики и служить, таким образом, базой для грамматик, словарей и другой надежной справочной литературы»[2] [Sinclair, 1996].

Идея о необходимости такого корпуса не нова. А.П. Ершов еще в 1978 г. выступил с инициативой создания «машинного фонда русского языка», а в некоторых странах электронные коллекции текстов собираются еще с конца 50-х гг. Так, начало базы FRANTEXT во Франции было положено в 1957 г., когда началась подготовка «Тезауруса французского языка» (*Trésor de la langue française*), основанного на корпусе текстов XIX-XX вв.

Разумеется, корпус, в полной мере отвечающий требованию репрезентативности, – это идеал, достичь которого едва ли возможно. Однако даже отдаленное к нему приближение дает в руки лингвистам (и не только лингвистам!) мощнейший инструмент исследования языка (а через язык – культуры народа).

4. Гуманитарная роль корпусов

2 «A reference corpus is one that is designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries thesauri and other language reference materials».

Весьма существенной представляется общегуманитарная роль больших текстовых корпусов. Можно сказать, что корпус – это новая своеобразная форма жизни языка. В отличие от бумажных картотек, которые после завершения исследования или издания, для которого они были предназначены, в лучшем случае попадают на хранение в архив, электронный корпус продолжает жить, обогащаться, сливаться с другими корпусами и активно служить последующим поколениям филологов. Разумеется, при условии, что этот корпус устроен так, что его можно интегрировать с другими корпусами, и очередная революция в технологии не сделает его малопригодным для дальнейшего использования.

Помимо решения собственно научных задач корпус текстов может использоваться в дидактических и даже чисто практических целях. Всем, кому приходилось писать тексты на неродном языке, известна проблема: даже самые хорошие словари с большим числом примеров не всегда позволяют сделать вывод, насколько «естественно» звучит та или иная конструкция и насколько точно она отражает вложенный в нее смысл. Хорошо, если «под рукой» есть носитель языка (сам к тому же обладающий хорошим чувством стиля). А теперь представьте себе, что у нас есть возможность проверить, встречается ли такая конструкция в корпусе текстов, и если да, то в каком контексте и в каких произведениях. К сожалению, осуществить такую мечту на практике пока невозможно. Технически это было бы несложно, но существующие большие корпуса текстов в настоящее время закрыты для свободного доступа.

Между тем открытость текстовых фондов является характерной чертой «идеологии» корпусной лингвистики. Как мне представляется, развитие корпусных исследований будет вести к формированию открытого лингвистического сообщества, свободно обменивающегося данными и активно развивающего совместные и взаимодополняющие исследовательские проекты. В этом отношении Машинный фонд русского языка (о котором речь пойдет ниже), предоставляющий через Интернет открытый доступ к своим ресурсам, идет в авангарде мировых тенденций. К сожалению, возможности, которые Машинный фонд предлагает исследователям, пока весьма скромны.

5. Инструменты работы с корпусом

Помимо собственно текстов полноценный корпус должен располагать набором «инструментов» для работы с ними. Инструменты эти можно подразделить на две категории: 1) средства просмотра текстов и запроса данных; 2) средства обогащения корпуса аналитической информацией, которая называется аннотацией (annotation), или разметкой (markup, tagging).

Наиболее распространенными способами просмотра текста являются имитация издания (с возможным выделением интересующих исследователя объектов) и конкордансы (список словоформ или словосочетаний в контексте). Основное преимущество электронного издания перед печатным состоит в возможности быстрого поиска интересующих исследователя форм и сочетаний. Широта параметров поиска зависит от того, какая аналитическая информация закодирована в корпусе.

Если мы хотим найти все случаи употребления определенной словоформы, то это легко сделать в простом текстовом файле. Если мы хотим найти все случаи употребления определенной лексемы, представленной рядом словоформ, это несколько сложнее, но также возможно. Если же мы хотим найти все случаи употребления определенной граммемы (например, творительный падеж единственного числа существительного), сделать это на неразмеченном корпусе крайне проблематично.

6. Что такое разметка корпуса?

Как уже отмечалось, **разметка** – это обогащение корпуса разного рода аналитической информацией.

Минимальная разметка, как правило, легко проводящаяся в автоматическом режиме, состоит в оснащении корпуса **ссылочной информацией**. Иными словами, когда мы получаем ответ корпуса на наш запрос, мы должны четко знать «координаты» нашего примера («текст / глава / абзац» или «страница / строчка»).

Для лингвистических исследований большую ценность имеет **морфологическая разметка**: каждая словоформа соотносится с «начальной» («словарной») формой лексемы, определяется ее частеречная принадлежность, граммемы словоизменительных категорий.

Для многих языков, в том числе русского, разработаны программы автоматической морфологической разметки, однако все они дают тот или иной процент брака (неизбежного по причине языковой омонимии) и требуют «ручной» проверки. В ряде случаев можно, однако, ограничиться грубыми автоматическими данными и учитывать процент погрешности.

Возможны и другие виды лингвистической разметки: синтаксическая, семантическая, прагматическая и т.п., однако их универсальность не столь очевидна. Если в вопросе об основных частях речи и составе граммем можно говорить об относительном консенсусе среди лингвистов, то синтаксические функции и семантические группировки разными языковедческими школами понимаются отнюдь не одинаково.

Таким образом, возникает отдельная методологическая проблема: как сделать так, чтобы различия и даже противоречия лингвистических теорий и исследовательских интересов не мешали успешному функционированию корпуса на благо всего лингвистического сообщества? Следует сразу отметить, что существуют технологии, позволяющие решить эту проблему.

Помимо лингвистической, существует и **филологическая** разметка. Она позволяет включать в корпус варианты текста, авторскую и редакторскую правку, выделять иностранные слова, цитаты, прямую речь персонажей литературного произведения, разного рода стилистические фигуры.

Аналитическая разметка корпуса – весьма трудоемкий процесс, однако он сам по себе не лишен научного интереса. В ходе «наклеивания ярлыков» на словоформы или синтаксические конструкции выявляются «узкие места» используемых классификаций, обращают на себя внимание интересные примеры. А главное, что результаты этой кропотливой работы не станут пылиться в архивах, а будут активно использоваться и развиваться.

7. Стандарты оформления корпусов

До сих пор мы говорили о свойствах корпуса в отвлечении от конкретных технологических решений, позволяющих воплотить их на практике. Такие решения могут быть различными, причем чем больше разметки содержит корпус, чем больше прикладных программ разрабатывается для его эксплуатации, тем более разнообразными и трудносовместимыми могут становиться технические решения.

Несовместимость стандартов, используемых создателями корпусов в разных странах и исследовательских центрах, ставит под угрозу столь важную для корпусной лингвистики возможность широкого обмена данными, объединения и взаимного обогащения корпусов.

7.1. Text Encoding Initiative (TEI)

В этих условиях с середины 80-х гг. начинается движение за принятие неких

международных норм оформления электронных изданий текстов. В 1987 г. это движение организационно оформилось и получило название «Инициатива по кодированию текстов» (Text Encoding Initiative, или TEI).

Основным продуктом деятельности TEI являются «Рекомендации по кодированию и обмену электронными текстами» с использованием стандартов SGML или XML. В настоящее время опубликована 4-я версия «Рекомендаций», занимающая в распечатанном виде более 1000 страниц. Следует, однако, подчеркнуть, что полная разметка корпуса текстов с использованием всех предлагаемых TEI элементов не только не является обязательной, но и вряд ли практически осуществима. Достоинство открытого формата S/XML[3] состоит как раз в том, что он позволяет осуществлять разметку частично и поэтапно: никакие «тэги»[4] за исключением минимальной разметки заголовка и «тела» текста не являются обязательными.

Что же представляют собой форматы SGML и XML?

7.2. Общая характеристика стандартов SGML и XML

Идея содержательной разметки электронных текстов (в отличие от кодировки формальных типографских средств) восходит к концу 1960-х гг. Впервые высказанная У. Танниклиффом и С. Райсом, она получает развитие в рамках исследовательского проекта компании IBM.

Разработка собственно формата SGML началась в 1978 г. под эгидой Американского национального института стандартизации (ANSI), позднее она была поддержана международным Институтом стандартизации (ISO). Название стандарта (Standard General Markup Language) можно перевести на русский язык как ‘стандартный обобщенный язык разметки’. Его первый черновой вариант был предложен в 1980 г., а в 1986 г. была официально опубликована окончательная версия (стандарт ISO 8879:1986)[5].

Стандарт SGML был использован для оформления документов ряда крупных издательств и промышленных корпораций, однако практика показала, что реализованный в нем принцип экономии разметки существенно осложняет последующую автоматизированную обработку документов. В середине 90-х гг. ведущие компании-производители программного обеспечения делегировали своих представителей в рабочую группу по созданию более удобного в эксплуатации «подвида» SGML, который получил название XML (Extensible Markup Language, ‘расширяемый язык разметки’). Одной из задач, которые ставились в начале разработки XML, было удобство использования этого формата для создания вебстраниц.

Сохраняя все основные достоинства SGML (за исключением некоторых возможностей сокращения «избыточной» разметки), XML значительно легче поддается обработке с целью визуализации и анализа данных. В числе наиболее популярных средств обработки документов XML следует назвать «стилевые листки» (stylesheets) в стандарте XSL. В настоящее время формат XML получает все более широкое распространение в сфере хранения данных, предназначенных для передачи через Интернет.

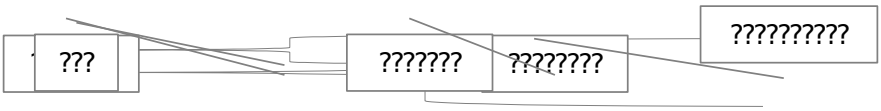
В основе S/XML[6] лежит представление о документе (будь то литературное

3 Мы будем пользоваться этим обозначением в тех случаях, когда разница между форматами SGML и XML не имеет значения.

4 Англ. *tag* ‘метка, ярлык, этикетка’ – термин, используемый для обозначения кодов разметки текстовых документов.

5 Более подробно история SGML представлена на сайте: <http://www.sgmlsource.com/history/sgmlhist.htm>.

6 Мы будем пользоваться этим обозначением, когда сказанное одинаково применимо к обоим стандартам: SGML и XML.



произведение, выпуск периодического издания или свод законов) как об иерархической структуре вложенных друг в друга *элементов*. Так, например, текст романа Л.Н. Толстого «Война и мир» состоит из четырех томов, каждый из которых включает несколько частей, каждая часть – глав, а каждая глава – абзацев. В тексте произведения могут встречаться отдельные слова или фразы на иностранном языке, иногда выделяемые шрифтом (курсивом). Эти фрагменты могут в электронной версии помечаться как особый элемент, хотя это не обязательно.

Каждый элемент помимо содержания («вложенного» элемента или текста) может иметь один или несколько *атрибутов*, содержащих дополнительную информацию – например, номер главы или иностранный язык, слово из которого использовано в тексте.

Правила того, какие элементы может содержать данный элемент и какие атрибуты он может иметь, записываются в специальный файл, который называется Document Type Declaration ‘декларация типа документа’ (DTD). В принципе для каждого документа можно создать собственную DTD, однако для удобства обмена текстовыми архивами целесообразно придерживаться по возможности единых «правил игры». Именно такую универсальную DTD стремится разработать TEI, а ее рекомендации можно рассматривать как комментарий к этой DTD.

Тэги S/XML представляют собой некоторый код, заключенный в угловые скобки. Код этот состоит из названия элемента (одна или несколько латинских букв) и атрибутов, которые отделяются пробелом и состоят из названия атрибута, знака равенства и значения в кавычках. Содержание элемента помещается между начальным и конечными тэгами (конечный тэг содержит только название элемента, перед которым ставится косая черта, или «слэш»). Ниже мы приводим фрагмент текста с выделенным с помощью тэгов иностранным словом:

```
Сперва <foreign lang="french">Madame</foreign>
за ним ходила
```

Теперь приведем пример оформления документа (в данном случае – романа А.С. Пушкина «Евгений Онегин») как иерархической структуры элементов: текст / глава / строка / стих. Названия элементов и их атрибуты соответствуют рекомендациям TEI.

```
<text>
  <div type=”chapter” n=”1”>
       <lg type=”stanza” n=”1”>
            <l>«Мой дядя самых честных
правил,</l>
            <l>Когда не в шутку занемог,</l>
            ...
       </lg>
       ...
  </div>
```

```
   ...
</text>
```

Более подробно общие принципы оформления документов в стандарте XML представлены во 2-й главе «Рекомендаций» TEI.

Как уже отмечалось в основе XML лежит идея содержательной разметки: тэги используются для кодирования не формы представления тех или иных элементов текста, а их функции в структуре текста. В этом отличие XML от HTML или кодов форматирования, используемых в текстовых процессорах (таких, как Microsoft Word[7]). Важнейшими особенностями формата S/XML являются его открытость (возможность поэтапной разметки, включения в нее дополнительных элементов) и отсутствие «командных кодов», «привязанных» к какому-либо определенному программному обеспечению: в принципе любой документ S/XML может читаться в любом простейшем текстовом редакторе.

Это дает основание надеяться, что использование названного формата обеспечит сохранность и доступность данных независимо от будущих изменений в программном обеспечении и научных теориях, на основании которых проводится разметка.

7.3. Проблемы, связанные с форматом XML, и их решения

Наряду с несомненными преимуществами формат XML имеет определенные неудобства.

Прежде всего проблемы возникают в связи с жесткой иерархической структурой составляющих документ элементов. В действительности, любой печатный или рукописный документ имеет как минимум две структуры – формальную (книга / страница / строка) и содержательную (произведение / часть / абзац / предложение / слово), которые не всегда совпадают. Например, предложение или отдельное слово могут начинаться на одной строке или странице и заканчиваться на другой. Однако элементы S/XML могут только «вкладываться» друг в друга, но не «пересекаться». Иными словами, мы не можем начать элемент «предложение» внутри одного элемента «страница», а закончить – внутри другого.

Согласно рекомендациям TEI, данная проблема может решаться двумя путями. Первый из них состоит в использовании «пустых элементов» или «меток» (milestone). Пустой элемент не имеет содержания, его начальный тэг одновременно является конечным. Так, к примеру, вместо элемента «страница», содержащего в себе текст, можно воспользоваться «меткой» «разрыва страницы», который можно поместить внутрь абзаца и даже слова (если слово содержит перенос со страницы на страницу). В XML такие элементы обозначаются тэгами с косой чертой перед конечной угловой скобкой (например, в начале третьей страницы может быть поставлена «метка» **<pb n="3"/>**).

Другое решение состоит в том, чтобы связать части «разорванного» элемента с помощью специальных атрибутов. Подобное решение может, к примеру, использоваться при разметке прямой речи, распространяющейся на несколько абзацев. Практическая реализация обоих предложенных решений связана с рядом сложностей и ограничений, поэтому при разработке концепции корпуса следует внимательно взвесить все «за» и «против» принятия той или иной системы разметки пересекающихся структур текста.

Еще одна проблема, связанная с использованием стандарта XML, – это его громоздкость. Чем более детальной становится разметка текста, тем сложнее оказывается его чтение с помощью простого текстового редактора. Если конечный

7 В определенном смысле «содержательное» форматирование в Microsoft Word все же возможно – через использование «стилей».

пользователь будет работать с текстом, представленным в оптимальной для него форме с помощью стилевых листков, то составители корпуса должны видеть непосредственно исходный код с тэгами.

Для облегчения и ускорения их работы можно воспользоваться двумя методами. Первый состоит в использовании упрощенного кода на стадии ввода данных. Например, можно использовать специальные символы или сокращенные обозначения вместо элементов XML на стадии ввода данных. Первичную разметку документа можно проводить в стандарте SGML, а затем воспользоваться специальными программами для автоматического преобразования SGML в XML. Данный метод следует использовать с большой осторожностью, так как он повышает риск ошибок при последующей обработке.

Другой способ, позволяющий облегчить редактирование «насыщенного разметкой» документа XML, – использование специального программного обеспечения («редакторов XML»). В настоящее время различные фирмы-разработчики предлагают целый ряд таких редакторов, каждый из которых имеет свои сильные и слабые стороны. Некоторые из них (особенно бесплатные) оказываются сложными в использовании для людей, не являющихся профессиональными программистами, другие стоят достаточно дорого и при этом предлагают множество функций, в которых составители корпусов не нуждаются. Таким образом, выбор оптимального программного обеспечения для работы с XML является отдельной немаловажной проблемой, которую следует решить в начале работы над проектом корпуса.

7.4. Сложности в реализации рекомендаций TEI

Несмотря на то, что рекомендации TEI составляются с целью учесть потребности всех возможных видов корпусов текстов и электронных изданий, на практике нередко обнаруживается, что эти рекомендации либо не содержат нужных элементов, либо свойства предлагаемых элементов не отвечают требованиям составителей корпуса. Следует отметить, что TEI постоянно ведет работу по усовершенствованию рекомендаций, и возможность отклонения от рекомендованной схемы не исключается самими разработчиками.

В качестве примера, когда рекомендуемого TEI набора элементов оказывается недостаточно, можно привести дипломатические транскрипции средневековых рукописей, с максимальной детализацией отражающие данные графической системы оригиналов (каллиграфические варианты букв, сокращения, пунктуацию, нестандартное разделение слов и т.п.). Опубликованные на сегодняшний день рекомендации TEI просто не содержат необходимых для адекватной кодировки этих объектов элементов[8].

Кроме того, заложенная в основу рекомендаций TEI концепция структуры текста в ряде случаев представляется дискуссионной. Так, TEI проводит фундаментальное различие между структурой прозаического и стихотворного текста. Базовой структурной единицей в прозе признается абзац (элемент **<p>**), а в стихотворном тексте – строка (элемент **<l>**). Более крупные единицы структуры стихотворного текста (четверостишия, строфы, куплеты и т.п.) обозначаются элементом **<lg>**. Подобное положение не вызывает возражений применительно к произведениям классической и современной литературы. Однако в средневековый период стихотворная форма широко использовалась в эпических и даже дидактических произведениях. Их структурные единицы (фрагменты, начинавшиеся с цветных буквиц) часто определялись исключительно содержанием и не имели определенных метрических характеристик. Для их обозначения термин

8 Следует, однако, отметить, что в настоящее время в рамках TEI действует рабочая группа по разработке системы кодировки транскрипций манускриптов.

«абзац» представляется более адекватным, чем «строфа», однако DTD, предлагаемая TEI, запрещает использовать элемент **<l>** в качестве содержания элемента **<p>**.

В подобной ситуации возможны различные решения. Во-первых, можно включить в DTD новые элементы, не предусмотренные TEI. Последние рекомендации консорциума предусматривают специальный модуль для таких элементов.

Во-вторых, можно внести изменения в декларированные свойства элементов TEI (например, разрешить использование абзацев в стихотворных текстах).

Наконец, можно, ничего не меняя в предложенной TEI DTD, использовать элементы или атрибуты в несвойственной им функции.

Наиболее разумным представляется первое решение, однако в тех случаях, когда расхождения между потребностями корпуса и возможностями, предоставляемыми стандартными рекомендациями TEI, невелики, третье решение может оказаться целесообразным. В любом случае все модификации в разметке корпуса по отношению к стандартной схеме TEI должны быть четко зафиксированы и мотивированы.

7.5. Лингвистическая разметка в XML-TEI

Рекомендации TEI не содержат полного списка атрибутов и их значений, которые следует использовать в лингвистической разметке. Это вполне объяснимо, поскольку многие вопросы классификации слов и грамматических категорий, не говоря уже о синтаксических конструкциях, остаются до настоящего времени дискуссионными.

Как уже отмечалось, наиболее востребованной в различных исследованиях и наименее зависимой от теоретических расхождений между различными лингвистическими школами может быть элементарная морфологическая разметка, т.е. соотнесение словоформы с «начальной» (словарной) формой лексемы, указание ее частеречной принадлежности, граммем классифицирующих и словоизменительных морфологических категорий.

Разумеется, и при такой разметке невозможно избежать спорных вопросов. Так, при работе с русскими текстами возникает вопрос о категории состояния, существование которой признается не всеми исследователями; об отдельных разрядах местоимений и числительных, которые нередко включаются в состав других частей речи. Некоторые ученые полагают, что качественные отадъективные наречия не являются самостоятельными лексемами [Панов, 1999].

При составлении протокола разметки следует четко эксплицировать сделанный в каждом спорном случае выбор. Желательно использовать как можно более детальную классификацию, т.к. при последующей эксплуатации корпуса значительно легче «нейтрализовать» излишне детальные оппозиции, чем разграничивать не различавшиеся при первоначальной классификации категории.

Наиболее адекватным методом морфологической разметки словоформ в формате XML является представление каждой словоформы в качестве отдельного элемента (**<w>**, согласно рекомендациям TEI) и использования атрибутов для записи аналитической информации. В качестве примера мы можем привести разметку словоформы *шутку* из второй строчки «Евгения Онегина»:

```
<w lemma="шутка" class="S-f-inan" form="sg-acc">шутку</w>
```

Сама словоформа здесь является содержанием элемента (она, в отличие от тэгов, не выделена жирным шрифтом). Внутри стартов тэга мы видим три атрибу-

та: **lemma**, **class** и **form**. Значением первого атрибута является начальная (словарная) форма лексемы. В значениях атрибутов **class** и **form** закодированы соответственно классифицирующие (S – существительное, f – женского рода, inan –неодушевленное) и словоизменительные (sg – единственное число, acc – винительный падеж) категории[9]. Некоторые другие примеры возможных способов лингвистической разметки и общие принципы, которых следует придерживаться, приводятся в Рекомендациях TEI (глава 15) и CES (часть 5).

Как мы уже отмечали, существуют программы автоматической морфологической разметки русских текстов, однако они используют «собственный» формат ее представления, что затрудняет эксплуатацию корпуса с использованием других программ и ставит его жизнеспособность в зависимость от наличия у пользователей соответствующих программ и их совместимости с операционными системами[10].

8. Российские и зарубежные корпуса текстов

В заключение настоящей статьи мы бы хотели привести несколько примеров существующих зарубежных корпусов текстов и кратко охарактеризовать современное состояние корпусной русистики (не претендуя, разумеется, на полноту охвата материала).

В целом можно разграничить два подхода к организации функционирования корпусов или коллекций текстов. Первый из них позволяет исследователю «скачивать» сами тексты (с разметкой или без нее) и затем работать с ними, используя как программы эксплуатации, рекомендованные составителями коллекции, так и свои собственные. Второй подход состоит в том, что исследователю предоставляется доступ к программе эксплуатации корпуса, но не к самим текстам. Последний подход, как правило, объясняется соображениями защиты авторских прав на тексты и их разметку.

Среди представленных ниже корпусов текстов первый подход реализован в рамках Машинного фонда русского языка, второй – в базе Франтекст и Национальном корпусе русского языка. Британский национальный корпус в известной мере совмещает оба подхода.

8.1. Британский национальный корпус (BNC)

Одним из наиболее известных и популярных корпусов английского языка (однако далеко не единственным) является Британский национальный корпус (British National Corpus, BNC). Этот корпус был создан совместными усилиями нескольких британских университетов и издательств[11], а также Британской библиотеки в период с 1991 по 1994 гг. Корпус включает письменные и устные тексты на британском английском конца XX в., принадлежащие к самым различным жанрам и функциональным стилям. Корпус является фрагментарным: тексты объемом более 45 000 слов представлены отрывками (что позволяет избежать влияния индивидуального стиля того или иного автора на общие результаты). Общий объем корпуса составляет

9 Условные обозначения частей речи и грамматических категорий позаимствованы из системы Национального корпуса русского языка <http://www.ruscorpora.ru/corpora-morph.html>.

10 Подробный анализ программ морфологической разметки и других средств эксплуатации корпусов русских текстов представлен в [Bénet, 2003].

11 В частности, Окфордским и Ланкастерским университетами, издательствами Oxford University Press и Addison-Wesley Longman.

немногим более 100 000 000 словоупотреблений. Тексты BNC размечены в стандарте SGML в соответствии с рекомендациями TEI.

Корпус BNC снабжен морфологической разметкой: каждая словоформа охарактеризована по принадлежности к части речи, разряду в рамках части речи и форме словоизменения. Эта разметка проводилась автоматически, что привело к ошибкам в 1,7% случаев, а 4,7% словоформ не смогли быть однозначно проинтерпретированы и получили «двойной морфологический код». Фрагмент корпуса, составляющий 2% от его общего объема, был отобран для более детальной («ручной») морфосинтаксической разметки.

Эксплуатация корпуса осуществляется с помощью ряда специально созданных программ обработки SGML. Ограниченный доступ к ресурсам корпуса бесплатно предоставляется через сеть Интернет <http://www.natcorp.ox.ac.uk/>, однако для того, чтобы воспользоваться всеми его возможностями, необходимо приобрести CD-ROM или за плату зарегистрироваться для доступа в режиме «on-line».

Данные BNC широко используются при составлении словарей, грамматик и учебников английского языка, в лингвистических исследованиях, в работах по искусственному интеллекту, а также в практике преподавания английского языка.

8.2. FRANTEXT

Одной из первых исторически и крупнейших на сегодняшний день электронных коллекций текстов является французская база Франтекст (Frantext). Строго говоря, это не корпус, однако система его эксплуатации позволяет исследователю формировать свой «рабочий корпус» с учетом целого ряда параметров (автор, дата, жанр, размер и др.).

Как уже отмечалась, работа по созданию базы началась в 1957 г. в рамках подготовки 16-томного «Тезауруса французского языка», однако со временем пополнение и разработка средств эксплуатации корпуса выделились в самостоятельную задачу. В создание Франтекста были вложены большие финансовые средства: целая лаборатория Национального центра научных исследований Франции (CNRS) в составе от 30 до 50 человек трудилась над ним в течение без малого полувека. В настоящее время Франтекст насчитывает 3737 текстов XVI – XX вв. (около 210 000 000 словоупотреблений) и продолжает непрерывно пополняться. Основную массу (около 80%) составляют литературные тексты, однако в ней также представлены научные и технические произведения. Немногим более половины текстов базы (1940 текстов, 127 000 000 словоупотреблений) снабжены морфосинтаксической разметкой.

Внешний доступ к Франтексту открыт с 1992 г. для корпоративных пользователей (библиотек, университетов и т.п.) и является платным. Свободный доступ предоставляется к библиографической базе и к электронной версии «Тезауруса французского языка» (TLFI).

В последние годы ведется работа по углублению «исторической перспективы» Франтекста: к нему добавлены базы текстов старофранцузского (IX – XIII вв.) и среднефранцузского (XIV – XV вв.) периодов, причем этими базами любой желающий может пользоваться бесплатно.

В какой-то мере достоинства Франтекста – его колоссальные размеры и длительная история формирования – являются в то же время источником его проблем. Разработанные в 60-е – 70-е гг. форматы и системы эксплуатации в настоящее время сильно устарели и не отвечают возможностям современной техники и запросам исследователей. Модернизация Франтекста – в частности, его перевод в стандарт XML и разметка в соответствии с рекомендациями TEI – является сложной задачей, и в настоящее время трудно сказать, когда она будет решена.

Тем не менее действующая система эксплуатации Франтекста позволяет ре-

шать многие лингвистические и литературоведческие задачи и широко используется исследователями французского языка во всем мире.

8.3. Корпусная русистика

Как мы уже отмечали, в нашей стране идея создания корпуса русских текстов, или «машинного фонда русского языка» принадлежит академику-информатику А.П. Ершову (1931 – 1988) и была высказана им в 1978 г. Практическая реализация проекта началась в середине 80-х гг. в Институте русского языка АН СССР. В 1986 г. был опубликован сборник работ, озаглавленный «Машинный фонд русского языка: идеи и суждения», а в 1989 г. вышла в свет монография В.М. Андрющенко «Концепция и архитектура Машинного фонда русского языка». В этих книгах подробно представлены все аспекты этого большого и чрезвычайно интересного проекта. Помимо обширного корпуса текстов Машинный фонд русского языка должен был включать словарно-грамматические базы данных на основе академических словарей и грамматик, фонд лингвистических алгоритмов и программ, включая процессоры русского языка.

К сожалению, экономический кризис начала 90-х гг. и связанное с ним катастрофическое сокращение финансирования научных исследований помешали реализации этих замыслов. С 1996 г. Машинный фонд русского языка развивается в основном за счет грантов Российского гуманитарного научного фонда (РГНФ) и Российского фонда фундаментальных исследований (РФФИ). В настоящее время отдел Машинного фонда поддерживает страницу в Интернете <http://www.irlras-cfrl.rema.ru/>, на которой представлены публикации сотрудников, несколько словарей, архив текстов русской литературы XIX-XX вв. (всего около 10 миллионов словоупотреблений), а также корпус газетных текстов конца 90-х гг. XX в. (представлены полные тексты 9 газет за вторую половину 1997 г., оформленные в стандарте SGML в соответствии с рекомендациями TEI). Сотрудниками отдела Машинного фонда (В.М. Андрющенко, Ж.Г. Аношкиной и Л.И. Колодяжной) были разработаны различные варианты системы Unilex, предназначенной для обработки и разметки текстов и ведения автоматических словарей. Общий обзор деятельности Машинного фонда русского языка представлен в отчете, опубликованном на его сайте в Интернете.

В 2003 г. группа лингвистов из Москвы, Санкт-Петербурга и ряда других научных центров приступила к работе над созданием «Национального корпуса русского языка». В перспективе, по замыслу участников проекта, корпус должен будет охватывать период с XIX до конца XX в. и измеряться сотнями миллионов словоупотреблений. Из информации на сайте проекта <http://www.ruscorpora.ru> пока неясно, координируется ли он с Машинным фондом русского языка. В настоящее время в корпусе представлены в основном литературные произведения конца XX вв. Тексты корпуса снабжены морфологической разметкой, разработанной под руководством В.А. Плунгяна на основе системы «Грамматического словаря» А.А. Зализняка. Эксплуатация корпуса осуществляется через «онлайновую» поисковую систему, функционирующую пока в тестовом режиме.

Значительно более крупные массивы текстового материала представлены в рамках так называемых «виртуальных библиотек». Наиболее крупной и известной из них является «Библиотека Мошкова» <http://www.lib.ru>. Общий объем представленных в ней текстов измеряется сотнями миллионов словоупотреблений. Следует, однако, подчеркнуть, что сами по себе электронные библиотеки не являются корпусами текстов, так как различные функциональные стили языка и жанры литературы представлены в них неравномерно. Значительную часть фондов виртуальных библиотек составляют переводные произведения. Последние могут, несомненно, служить интересным источником материала (в особенности для

сопоставительных исследований), однако их следует четко отграничивать от «исконно русских» текстов и тщательно исследовать на предмет возможных калек. В ряде случаев тексты электронных библиотек содержат значительное число опечаток (ошибок распознавания при сканировании). Наконец, в отечественных электронных библиотеках не всегда урегулирован вопрос авторских прав на публикуемые произведения. Тем не менее виртуальные библиотеки с их богатством и разнообразием текстового материала потенциально представляют собой ценнейший ресурс для лингвистических исследований.

В последние годы, особенно в студенческих курсовых и дипломных работах, достаточно часто фигурируют примеры, источник которых определяется как «Интернет». Подобная практика недопустима в научном исследовании, так как сама по себе «всемирная паутина» может рассматриваться в качестве корпуса текстов в еще меньшей степени, чем виртуальные библиотеки. Без четкого указания на источник примера и определения его функционально-стилистической и жанровой принадлежности (при том что, насколько нам известно, «жанровая классификация» текстов Интернета еще не разработана) невозможно оценить языковой статус иллюстрируемого примером факта. Вместе с тем Интернет безусловно представляет собой новую и чрезвычайно интересную среду бытования языка со свойственными ей уникальными жанрами («чатами», «форумами», электронной перепиской, записями в гостевых книгах и т.п.), которая заслуживает самого пристального внимания лингвистов.

Заключение

Создание и развитие самых разнообразных корпусов текстов на разных языках – как «больших», так и находящихся под угрозой исчезновения – с полным основанием можно признать одной из приоритетных задач лингвистики. Эти корпуса дадут в руки будущих поколений исследователей надежный и легкодоступный источник данных о функционировании языка в самых различных сферах и о культуре народа, говорящего на этом языке. При создании корпусов текстов следует ориентироваться на международные стандарты и рекомендации, призванные обеспечить сохранность и доступность данных независимо от изменения технологий и программного обеспечения.

Ресурсы в Интернете

Британский национальный корпус: <http://www.natcorp.ox.ac.uk/>

Корпус FRANTEXT: <http://www.atilf.fr>

Краткая история развития SGML: <http://www.sgmlsource.com/history/sgmlhist.htm>

Машинный фонд русского языка: <http://www.irlras-cfrl.rema.ru/>

Русский национальный корпус: <http://www.ruscorpora.ru>

World Wide Web Consortium (спецификация XML): <http://www.w3c.org>

Text Encoding Initiative (TEI): <http://www.tei-c.org>

Corpus Encoding Standard (CES): <http://www.cs.vassar.edu/CES>

Литература

Андрющенко В.М. Концепция и архитектура Машинного фонда русского языка, М., 1989.

Машинный фонд русского языка: идеи и суждения. М., 1986.

Панов М.В. Позиционная морфология русского языка. М., 1999.

Bénet V. L'outil informatique au service du russiste français. Traitement informatique du russe. Thèse de Doctorat de 3ème cycle. Clermont-Ferrand, 2003.

Biber D., Conrad S., Reppen R. Corpus Linguistics: Investigating Language Structure and Use. Cambridge, 1998.

Sinclair J. Preliminary recommendations on Corpus Typology. Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards). 1996 <http://www.ilc.cnr.it/EAGLES96/pub/eagles/corpora/corpustyp.ps.gz>