О. Н. Ляшевская

Корпусные инструменты в грамматических исследованиях русского языка

ББК 81.1
УДК 80/81
Л 99

Издание осуществлено при финансовой поддержке
Фонда фундаментальных лингвистических исследований
проект № В-28-2014

Утверждено к печати Ученым советом
Института русского языка имени В. В. Виноградова РАН

Рецензенты:
д. ф-м. н. М. Р. Пентус, к. филол. н. И. В. Азарова

Ляшевская О. Н.

Л 29 Корпусные инструменты в грамматических исследованиях русского языка. — М.: Издательский Дом ЯСК: Рукописные памятники Древней Руси, 2016. — 520 с.

ISBN 978-5-9907947-8-8

Русская корпусная лингвистика представлена в книге двумя направлениями. Первая часть содержит описание подходов и методов аннотации Национального корпуса русского языка (http://ruscorpora.ru), включая разметку лексико-грамматической, лексико-семантической, семантико-синтаксической и словообразовательной информации. Кроме того, описываются процедуры оценки инструментов автоматической разметки текстов (морфологических и синтаксических парсеров) и идеология создания двух частотных корпусных словарей, общего и лексико-грамматического. Во второй части представлены результаты исследований грамматики и лексики русского языка с применением квантитативных корпусных методов: изучение грамматических, конструкционных и семантических профилей языковых единиц, в том числе глаголов и глагольных приставок, имен существительных и пространственных конструкций.

УДК 80/81
ББК 81.1

В оформлении переплета использована картина
Пита Мондриана «Серое дерево», 1911

ISBN 978-5-9907947-8-8

СОДЕРЖАНИЕ

Приложения

1.1. Национальный корпус русского языка и его аннотация

Принципам составления, разметки и использования представительных корпусов языков мира посвящена уже довольно объемная коллекция литературы, см. (O'Keeffe, McCarthy 2010; McEnery, Hardie 2012; McEnery, Wilson 2001; Tognini-Bonelli 2001; Захаров, Богданова 2011; Большакова и др. 2011); статьи журнала International Journal of Corpus Linguistics, материалы конференций «Corpus Linguistics», LREC, COLING и т. п., тематические сборники статей в ведущих издательствах мира, онлайн-курсы по корпусной лингвистике, профессиональная email-рассылка Corpora List и мн. др. Документацию по Национальному корпусу русского языка можно найти в сборниках (НКРЯ 2003—2005; НКРЯ 2006—2008; НКРЯ 2012—2014), в публикациях конференций «Диалог», MegaLing, CORPORA, «Манускрипт» и т. д. (многие публикации доступны на сайте корпуса http://ruscorpora.ru и на обучающем портале http://studiorum.ruscorpora.ru). Очень коротко, схема создания корпуса выглядит следующим образом:

- собрать и технически подготовить электронные версии текстов (в соответствии с заранее продуманным планом объема, временно́го и жанрово-тематического баланса текстовой коллекции);
- расклассифицировать тексты по сфере употребления, жанру, тематике, авторству, времени создания, источнику происхождения и т. п. и приписать соответствующий набор условных ярлыков-тегов каждому тексту (мета-текстовая аннотация);
- каждому слову текста приписать набор тегов частеречной принадлежности, леммы (словарной формы, начальной формы слова), других словоизменительных признаков (лексико-грамматическая аннотация);
- каждому предложению, отдельным словам, группам и составляющим приписать сведения о синтаксическом типе языковой единицы и типе синтаксического отношения между элементами (синтаксическая аннотация);

и т. п. — каждому языковому уровню, как правило, соответствует свой уровень аннотации в корпусе, начиная от кодирования фонетических цепочек и знаков препинания и заканчивая аннотацией дискурсивных стратегий и референциальных отношений. Иными словами, корпус — это коллекция текстов, в которую «воткан» длинный шлейф лингвистических знаний о каждой большой и малой единице языковой структуры.

Остается занести в базу данных координаты каждого аннотированного элемента, создать индексы для быстрого поиска, подключить словари для расширения возможностей поиска, загрузить все данные в специальную программу (корпус-менеджер, желательно работающий онлайн) и... корпусом можно пользоваться как информационно-справочной системой.

В качестве примера на рис. 1 приведено XML-представление разметки очень короткого фрагмента текста, где на три словоформы *Цены в них* приходится 79 строк разметки (и это не считая метаразметки, касающейся всего текста). Данный пример будет выдан, в числе прочих, поисковой системой корпуса, если пользователь задаст какой-либо признак (или комбинацию признаков) из тех, что содержатся в корпусной разметке.

В зависимости от типа исходного текста (включая звучащие источники в виде аудио- или видеофайлов, старые газеты, рваные объявления на заборе и т. п.), объема корпуса и задач, для которых он создается, будут различаться технологии первичной подготовки, количество уровней аннотации и детализированность системы тегов на каждом уровне, технологии самой разметки. Например, медиафайлы корпуса кинофильмов понадобится очистить от шумов, разрезать на короткие клипы, разметить временны́е границы реплик, сделать транскрипт звучащей речи, произвести разметку транскрипта как письменного текста, добавить разметку ударений, интонации, жестикуляции и мимики говорящего и т. п. В корпус древних документов имеет смысл добавить уровень представления графического вида слов и строк в рукописи, «перевод» на современный язык и, возможно, даже комментарии исследователей относительно возможных вариантов интерпретации текста. Кстати, небольшую коллекцию древних документов можно разметить вручную — тогда как для аннотации 100-миллионного корпуса новостей понадобится автоматическая программа.

Слово «технология» мы упоминаем не случайно: разметка корпуса — это всегда компромисс между наличием доступных компьютерных программ, электронных словарей, списков слов и других структурированных источников лингвистических данных, временем разметки и стоимостью оплаты труда разметчиков, а также требуемым качеством разметки в смысле полноты и точности.

О полноте и точности разметки требуется сказать отдельно. Для разных уровней аннотации полнота определяется по-своему, но в целом имеется в виду два понимания: количество элементов корпуса (слов, предложений, жестов и т. п.), охваченных аннотацией, и количество признаков и противопоставлений, учитываемых уровнем аннотации. Так, например, в корпусе может быть размечена морфемная структура всех слов *vs.* только самых частотных (сплошная — выборочная аннотация); все типы синтаксических отношений *vs.* синтаксические отношения, связывающие только предикат и его зависимые (богатая аннотация — бедная аннотация).

```
<word text="Цены">
 <ana>
  <el name="lex">
   <el-group>
    <el-atom>цена</el-atom>
   </el-group>
  </el>
  <el name="gramm">
   <el-group>
    <el-atom>S</el-atom>
    <el-atom>inan</el-atom>
    <el-atom>f</el-atom>
    <el-atom>pl</el-atom>
    <el-atom>nom</el-atom>
   </el-group>
  </el>
 </ana>
 <ana>
  <el name="sem">
   <el-group>
    <el-atom>r:abstr</el-atom>
    <el-atom>t:param</el-atom>
   </el-group>
  </el>
 </ana>
 <ana>
  <el name="flags">
   <el-group>
    <el-atom>animred</el-atom>
    <el-atom>capital</el-atom>
    <el-atom>first</el-atom>
    <el-atom>numred</el-atom>
    <el-atom>posred</el-atom>
   </el-group>
  </el>
 </ana>
</word>
<text> </text>
```

```
<word text="в">
 <ana>
  <el name="lex">
   <el-group>
    <el-atom>в</el-atom>
   </el-group>
  </el>
  <el name="gramm">
   <el-group>
    <el-atom>PR</el-atom>
   </el-group>
  </el>
 </ana>
 <ana/>
</word>
<text> </text>
```

```
<word text="них">
 <ana>
  <el name="lex">
   <el-group>
    <el-atom>они</el-atom>
   </el-group>
  </el>
  <el name="gramm">
   <el-group>
    <el-atom>SPRO</el-atom>
    <el-atom>3p</el-atom>
    <el-atom>pl</el-atom>
    <el-atom>loc</el-atom>
   </el-group>
  </el>
 </ana>
 <ana>
  <el name="sem">
   <el-group>
    <el-atom>r:pers</el-atom>
   </el-group>
  </el>
 </ana>
 <ana/>
</word>
<text> </text>
```

Рис. 1. XML-представление аннотации фрагмента текста НКРЯ: начало предложения *Цены в них ниже, чем в обычных магазинах*[1]

[1] В аннотации представлены лексико-грамматический (теги *lex* и *gramm*) и лексико-семантический (тег *sem*) уровни аннотации, а также уровень дополнительных «флагов». Полный список значений помет содержится на странице http://ruscorpora.ru. Под тегами *word* и *lex* приводятся орфографический вид словоформы и лемма соответственно. Далее, в данном примере комбинация S, inan, f, pl, nom обозначает неодушевленное существительное женского рода в форме им. падежа мн. числа (*цены*); PR — предлог (*в*); SPRO, 3p, pl, loc (*них*) — местоимение 3 лица в форме предл. падежа (*них*). Информация о лексико-семантических разрядах и группах, к которым относятся слова, кодируется тегами r:abstr, t:param (абстрактное параметрическое имя) и r:pers (личное местоимение). Флаги *capital* и *first* обозначают первое слово в предложении, написанное с заглавной буквы; *posred*, *animred*, *numred* указывают, что в слове повторяются значения признаков части речи,

Неточность разметки происходит в первую очередь из омонимии (неоднозначности), свойственной языку на самых разных уровнях. В приведенном примере аннотации (рис. 1) словоформе *цены* теоретически можно приписать две взаимоисключающие пары тегов — *gen sg* (род. падеж ед. числа) и *nom pl* (им. падеж мн. числа)[2], а словоформе *них* — взаимоисключающие теги *gen*, *acc* и *loc* (род., вин. и предл. падеж). Это омонимия на уровне словоизменения (грамматическая омонимия). Местоимение *них* может быть размечено как кореферентное одному из ранее упомянутых существительных, на выбор: *супермаркет*, *костел* и *страна* — это омонимия на уровне аннотации анафоры и кореференции. Глагол *загнуть* может быть аннотирован как глагол каузации изменения положения в пространстве (ср. *загнуть палец*) и глагол интерпретации речи (ср. *Ну ты загнул, брат!*) — это омонимия на лексико-семантическом уровне[3] и т. п.

В корпусной лингвистике омонимию технически определяют как альтернативные комбинации тегов разметки, которые можно приписать языковой единице, если не знать контекста ее употребления. Разрешение омонимии — это выбор наиболее подходящего варианта, исходя из контекста. Эта задача может быть поручена либо аннотатору-человеку, либо компьютерной программе. Компьютерная программа принимает решение, руководствуясь правилами, созданными лингвистами, или основываясь на статистической вероятностной модели. Например, правило выбора грамматических характеристик слова *Цены* может быть таким: «По умолчанию слово в начале предложения <начинающееся с заглавной буквы> стоит в именительном падеже»). Статистическая вероятностная модель сама предлагает множество подобных правил, в этом случае используется машинное обучение на ранее размеченной человеком части корпуса.

По точности разрешения омонимии компьютерные программы (пока еще) значительно уступают человеку, однако аннотатор не может быстро обработать миллионы контекстов в корпусе и, как замечено, в 3—5 % случаев все равно делает ошибки — по невнимательности, из-за недостатка лингвистической компетенции или недостаточной последовательности в принятии сложных решений. Производительность и последовательность может быть существенным фактором и для выбора порога точности в компьютерных приложениях. Простые, но менее точные алгоритмы могут оказываться более предпочтительными для обработки

одушевленности и числа предыдущего слова (в данном случае последнего слова предшествующего предложения).

[2] Поскольку ударения в электронной версии исходного текста не проставлены, статус омографов (*ценЫ* и *цЕны*) такой же, как и статус других омоформ, ср. *лечу* как форма глаголов *лечить* и *лететь*.

[3] Заметим, что полисемия и омонимия в корпусной аннотации обычно не противопоставляются. Таким образом, варианты семантических тегов для полисемичного глагола *загнуть*, для разных пониманий приставочного глагола *запустить* (ср. ‘каузировать летать’ и ‘привести в неудовлетворительное состояние’), для «чистых» омонимов типа *лук* (ср. ‘растение’, ‘оружие’, новое ‘фотография’) ничем не отличаются по статусу.

больших массивов корпусных данных. И наконец, заметим, что в целом далеко не всегда очевидно, что разрешенная омонимия — это абсолютное благо. Скажем, поиск в корпусе глаголов деформации и изменения пространственного положения (ср. *загнуть*) в роли глаголов речи, т. е. поиск с учетом «генетического» фактора или «внутренней структуры», — вполне осмысленная лингвистическая задача.

Потребности потенциального пользователя корпуса — это, пожалуй, самое важное, что влияет на содержание аннотации корпуса. Различают корпусы, созданные исследователями для себя и под свои конкретные исследовательские нужды (например, материалы фольклорных исследований или полевых экспедиций в малые языки), и общепользовательские корпусы, которые рассчитаны на многообразные нужды ученых, студентов, преподавателей языка и т. д. Национальные корпусы относятся ко второму типу. Если при разметке корпуса «для себя» исследователь может вводить какие угодно и очевидные только ему пометы, то разметка больших общепользовательских корпусов предполагает соблюдение ряда принципов:

- «очевидность» принятых помет и системы их противопоставления;
- наличие стандарта принятия решений при разметке данных.

Идеально, чтобы система используемых признаков была общепринята в сообществе потенциальных пользователей, например известна из стандартного школьного / университетского курса или описана в общепризнанной академической грамматике. Если признаки полагаются неизвестными «рядовому» пользователю, они должны быть просты для усвоения. В практике создания национальных корпусов обычно комбинируют бóльшую часть общеизвестных, традиционных помет с небольшим количеством помет, которые пользователь может освоить в короткое время.

Стандартная инструкция по разметке данных на том или ином уровне важна потому, что обычно эта задача поручается команде аннотаторов. Соответственно, они должны использовать одну и ту же систему помет и в идеальном случае принимать одинаковые решения в похожих типах контекстов. Стандарт аннотации включает описание принципов аннотации, наиболее характерные и сложные случаи использования тегов, а также сам тагсет — классификацию помет, желательно со статистикой их встречаемости в уже размеченной части корпуса.

Далее в этой части книги мы расскажем о нескольких проектах разметки корпуса, в которых принимал участие автор. Во второй главе речь пойдет о лексико-грамматической разметке, т. е. определении леммы, части речи и характеристик словоизменения словоформ. Глава охватывает задачи создания морфологического стандарта, создания ресурсов для разметки (электронного грамматического словаря) и проведения экспертизы качества работы компьютерных приложения.

Во третьей главе мы обратимся к лексико-грамматической разметке. Речь пойдет о принципах классификации лексики по группам типа «имена инструментов», «глаголы речи», «прилагательные цвета» и т. п., а также об экспериментах по разрешению лексической неоднозначности в контексте.

Четвертая глава посвящена разметке синтаксических и семантических отношений между элементами предложения, в частности о реализации в тексте лексических конструкций глагола (моделей управления и фразем). Описаны принципы создания ресурса ФреймБанк, основанного на данных Национального корпуса русского языка, а также представлен опыт оценки качества работы синтаксических парсеров.

В пятой главе мы обращаемся к представлению словообразовательной информации в корпусе.

Шестая глава описывает опыт создания частотных словарей на базе корпуса.

1.2. Словоизменение

1.2.1. Морфологический стандарт корпуса *

Эта глава посвящена теоретическим и практическим вопросам представления морфологической информации в корпусе текстов современного русского языка (вторая половина XX — начало XXI в.). Основой унифицированной аннотации языковых единиц является морфологический стандарт корпуса — совокупность решений, связанных со структурой морфологических категорий, с составом парадигмы слова и с единообразной трактовкой спорных вопросов русской грамматики. Эти решения должны, с одной стороны, учитывать грамматическую традицию и быть понятными для пользователей корпуса, а с другой стороны, должны допускать возможность практической реализации в технологическом процессе разметки.

Существующий опыт теоретического обсуждения и практического создания морфологически размеченных корпусов показывает, что можно выделить две крайности в подходах к аннотированию языковых единиц. Первый подход, который можно назвать формально-морфологическим, предполагает, что каждой встреченной в тексте словоформе, отличающейся по внешнему виду от других словоформ, присваивается некоторый ярлык вне зависимости от реально стоящей за ней грамматико-семантической или синтактико-семантической информации. Например, русской словоформе *брата* всегда приписывается ярлык «родительный падеж», даже если в некотором контексте эта словоформа с точки зрения «школьной» грамматики интерпретируется как винительный падеж: *Я привел своего брата.* То же касается информации о лексемной принадлежности словоформы: у омонимичных словоформ типа *были* (от глагола *быть*) и *были* (от существительного *быль*) исходной формой всегда будет считаться инфинитив глагола *быть*.

Второй подход, который можно назвать углубленным семантическим, нацелен на извлечение как можно более полной семантической информации, связанной с данной словоформой. Примером ярлыков в корпусе, размеченном согласно такой

---

* Первоначальный вариант текста опубликован в виде статей: *Ляшевская О. Н., Плунгян В. А., Сичинава Д. В.* О морфологическом стандарте Корпуса современного русского языка (Ляшевская и др. 2005б); *Ляшевская О. Н., Плунгян В. А., Сичинава Д. В.* О морфологическом стандарте Корпуса современного русского языка (Ляшевская и др. 2005а).

идеологии, могли бы служить пометы «настоящее историческое время» (для словоформ *приходит* и *смотрит* во фразе *А он вчера приходит и смотрит как-то странно*) или «будущее в значении вежливого побуждения» (для словоформы *передадите* во фразе *Не передадите ли вы мне соли?*).

Формально-морфологический подход часто применяется в прикладной лингвистике — в особенности в системах, где используется сплошное автоматическое аннотирование текстов. Он выгоден тем, что позволяет разметить огромные массивы текстов без участия человека (программа приписывает информацию, руководствуясь электронными морфологическими словарями-указателями словоформ). Кроме того, он прост (для установления морфологических характеристик программе не требуется анализировать контекст), удобен для статистических исследований, а отсутствие морфологической омонимии в разметке (т. е. ситуации, когда одной словоформе приписывается несколько конкурирующих морфологических разборов) позволяет избежать «комбинаторного взрыва» при автоматическом построении различных синтаксических и семантических гипотез.

Главный недостаток чисто морфологического подхода становится очевиден, если размеченный таким способом корпус предлагается пользователю-человеку (будь то лингвист, школьник, иностранец, изучающий русский язык и т. п.). Неподготовленный пользователь будет, по-видимому, весьма озадачен, получив по запросу «винительный падеж» только формы единственного числа женского рода на *-у* / *-ю* или узнав, что в русском языке родительный падеж употребляется после предлога *за* (ср. *Рад* ***за брата***). Поскольку формально-морфологический подход предлагает совершенно нестандартный взгляд на грамматику русского языка, идущий вразрез со сложившейся грамматической традицией, размеченный таким образом корпус будет малопригоден для использования в качестве экспертной системы по русскому языку.

С другой стороны, разметка текста в соответствии с углубленным семантическим подходом предполагает кропотливую работу лингвиста-эксперта, который анализирует особенности контекста, интонационные характеристики высказывания и т. п. К сожалению, пока не существует компьютерных программ, которые были бы способны заменить человека на этом направлении и обеспечить должный уровень адекватности, а значит, нереально обработать таким образом значительные объемы текстов. Вместе с тем стремление к максимальной детализации грамматического значения таит и иную опасность. Разметка субъективна, поскольку зависит от интуиции эксперта, и, следовательно, повышается вероятность, что другой носитель русского языка (или другой специалист) окажется не согласен с предлагаемой трактовкой грамматического значения словоформы.

Таким образом, каждая из представленных крайних точек зрения имеет свои достоинства и недостатки. В связи с этим идеальным балансом между ними кажется такой подход к морфологической разметке текста, при котором словоформы размечаются на уровне традиционных грамматических ярлыков, таких как «родительный падеж» или «настоящее время», а омонимичным словоформам приписывается

только одна и «правильная» (т. е. общепринятая в русской грамматической традиции) характеристика. Именно такой взгляд на устройство морфологической разметки сформировался в коллективе разработчиков корпуса, см. (Герд, Захаров 2004). Предполагается, что глубина семантической информации о грамматических формах достаточна для большинства пользователей корпуса[1], а задача выбора нужного значения в принципе алгоритмизуема; таким образом, морфологическая разметка больших по размеру корпусов может быть осуществлена, по крайней мере в значительной части, при помощи компьютера.

Однако информация о потенциальной грамматической многозначности словоформы, т. е. о морфологической омонимии, также не бессмысленна. Два вида размеченных текстов — один со снятой омонимией и другой, в котором омонимичным словоформам приписаны все возможные морфологические разборы, — могут быть полезны не только для тренировки «обучаемых» прикладных программ, но и для лингвистов, задавшихся вопросом: почему человек «не замечает» морфологической омонимии в тексте, например почему он не понимает форму *мыла* во фразе *Мама мыла раму* как форму родительного падежа существительного *мыло*?

Корпус современного русского языка (вторая половина XX — начало XXI в.) входит в Основной корпус НКРЯ и состоит из двух подкорпусов — со снятой и с неснятой грамматической омонимией. Разметка корпуса с неснятой омонимией осуществляется автоматически, тогда как разметка корпуса со снятой омонимией в настоящее время происходит в полуавтоматическом режиме (см. ниже) и требует участия человека. В связи с этим корпус с неснятой грамматической омонимией существенно превышает по размеру корпус со снятой грамматической омонимией. В поисковой системе, расположенной на сайте ruscorpora.ru, пользователь может задать ограничение на поиск по корпусу только со снятой или только с неснятой грамматической омонимией. Поиск по корпусу с неснятой омонимией дает гораздо больше языкового материала, но, поскольку омонимичные формы в нем получают весь возможный набор разборов, поисковая выдача по этим текстам содержит значительное количество «шума». Однако необходимо понимать, что разборы в корпусе с неснятой грамматической омонимией не являются ошибочными — они имеют другой статус: статус гипотетических разборов.

В следующих разделах мы представим технологию морфологической разметки, применяемую в корпусе[2], а затем обсудим особенности трактовки отдельных грамматических категорий и форм.

---

[1] Исследователь семантики грамматических категорий сможет сам провести необходимую детализацию значения, выбрав из предоставленного материала, например, по употреблениям форм настоящего времени, примеры на «обычное» настоящее и настоящее историческое. Скорее всего, разные исследователи сделают это несколько по-разному.

[2] Морфологический стандарт, разработанный для текстов Основного корпуса, используется также при разметке текстов газетного, устного, поэтического, мультимедийного, акцентологического корпусов и русской части параллельных корпусов. В разметке текстов

Морфологическая разметка в корпусе современного русского языка

Морфологическая разметка текста состоит в выделении словоформ и в приписывании каждой словоформе информации о лексемной принадлежности (исходной форме слова) и о совокупности ее грамматических признаков.

В результате морфологической разметки в тексте выделяется несколько видов текстовых фрагментов:

- русские словоформы (в том числе неопознанные и гипотетические словоформы), состоящие из букв кириллицы и, в редком случае, из знаков дефиса (-) и апострофа (’): *человек*, *что-то*, *д’Артаньян*;
- арабские или римские цифры, а также словоформы, основанные на цифровой основе, т. е. состоящие из арабских или римских цифр с добавлением букв кириллицы (часто также знака дефиса): *17*, *XIX*, *17-й*, *100-рублевый*;
- иноязычные фрагменты текста из словоформ, записанных латинскими, греческими и другими некириллическими буквами (*How do you do*, π), или из кириллических словоформ, представляющих запись текста на иностранном языке (*Гуд ивнинг*, *Здоровеньки булы*)[3];
- знаки препинания: точка, запятая, тире, кавычки, вопросительный, восклицательный знак, двоеточие, многоточие и нек. др.;
- прочие символы типа %, >, $ и др.

Все фрагменты текста, кроме русских словоформ, а в корпусе со снятой грамматической омонимией — еще и цифр и словоформ на цифровой основе (для них используется особая помета ciph), считаются неанализируемыми цепочками символов.

Морфологическая разметка содержит информацию о словоизменительных, но не о словообразовательных признаках лексемы. Информация о морфемном составе лексем представлена в слое словообразовательной разметки (см. главу 1.5). Деривационно-семантические признаки, такие как «диминутив», «имя деятеля», «сингулятив», «семельфактив», включены в состав лексико-семантической разметки, представляющей собой расширение морфологической аннотации (см. главу 1.3.1).

XVIII в. и обучающего корпуса используются различные расширения данного стандарта. Синтаксический, диалектный и исторические корпуса используют собственные стандарты морфологической разметки. Например, в синтаксическом корпусе представлена другая система показателей времени глагола, а в корпусе древнерусского языка аннотированы аналитические формы (да и сама структура грамматических тегов там, естественно, настроена на грамматическую систему древнерусского периода).

[3] Словоформы, записанные смесью кириллических, латинских и прочих символов (*e-mail’ы*, *PRuть* и т. п.), приравниваются к кириллическим, так как кириллические элементы в их написании говорят чаще всего в пользу адаптации недавних заимствований к грамматической системе русского языка и о появлении у них словоизменения.

Совокупность морфологических признаков, приписываемых словоформе в некотором значении, называется ее морфологическим разбором. Если какая-либо словоформа отождествляется с несколькими грамматическими значениями (наборами грамматических признаков), то ей изначально приписываются все возможные разборы. Используемые в морфологической разметке словоизменительные признаки мы будем называть также грамматическими признаками, а морфологические разборы — грамматическими разборами.

Морфологическая информация, приписываемая произвольному слову в тексте, состоит из четырех групп помет:

1. Лексема, которой принадлежит словоформа (указывается «словарная запись» данной лексемы, т. е. лемма).
2. Множество грамматических признаков данной лексемы, или словоклассифицирующие характеристики (указываются принадлежность лексемы к той или иной части речи и признаки, например, рода для существительного, переходности для глагола и т. п., а также сведения о несклоняемости имен существительных и прилагательных)[4].
3. Множество грамматических признаков данной словоформы, или словоизменительные характеристики (например, падеж для существительного, число для глагола).
4. Информация о нестандартности грамматической формы и орфографических особенностях написания словоформы (грамматически аномальные формы, орфографические искажения, аббревиация типа *млн*, *г-н* и т. п., написание с заглавной буквы, через дефис, цифровая запись).

Пометы первого, второго и третьего типа записываются в конкретный грамматический разбор, пометы четвертого типа приписываются словоформе в целом[5].

Морфологическую разметку дополняет так называемая акцентуационная разметка, в которой представлена информация о некоторых особенностях плана выражения словоформы, таких как место ударения и произношение *е* как «ё»[6].

В основу метаязыка грамматических помет, ввиду предполагаемой широкой международной аудитории пользователей корпуса, положена система сокращенных помет («тегов») на основе латинского алфавита. В то же время предусмотрена возможность использования при поиске традиционных названий категорий на русском

---

[4] В этой же зоне записываются пометы «фамилия», «имя», «отчество», «зооним» и «инициал», не являющиеся в строгом понимании словоклассифицирующими грамматическими характеристиками, но коррелирующие с типом словоизменения лексемы.

[5] Поиск по словоформе и лемме доступен в окне «Слово», по словоклассифицирующим и словоизменительным признакам — в окне «Грамм. признаки», а поиск по нестандартным пометам — в окне «Доп. признаки» лексико-грамматического поиска НКРЯ.

[6] Акцентуационная разметка не применяется в корпусе с неснятой омонимией, т. к. у омонимичных словоформ может быть несколько вариантов представления, ср. *больша́я* и *бо́льшая*, *лет* и *лёт*.

языке (в форме «грамматические признаки»). Полный список граммем и их сокращенную латинскую нотацию см. в разделе «Морфология» на сайте *ruscorpora.ru*.

Приведем пример разбора фразы *Вы оста-авите!*[7]:

<w><ana lex="вы" gr="SPRO pl 2p=nom"/>Вы</w>
<w><ana lex="оставить" gr="V pf tran=act fut 2p pl=distort"/>оста-авите</w>!
[Александр Солженицын. В круге первом (т. 1)].

Пример разбора словоформы со смешанным латинско-кириллическим написанием:

<w><ana lex="Ablaut" gr="S m inan=sg dat"/>Ablaut'y</w>.

Как уже было сказано, тексты корпуса размечаются автоматически (по крайней мере, на первом этапе) с помощью специальных программ — морфологических анализаторов. При разметке используются встроенные в эти программы морфологические словари, основанные на «Грамматическом словаре русского языка» А. А. Зализняка (Зализняк 1977/2003). Словари включают имена собственные, аббревиатуры типа *ЦСКА* и продуктивные части сложных слов типа *авто-*, *радио-*.

Разметка корпуса с неснятой лексико-грамматической омонимией осуществляется:

- автоматическим морфологическим анализатором, порождающим все потенциально возможные разборы словоформ, а также гипотезы относительно словоформ, отсутствующих в словаре[8];
- автоматическими фильтрами, поправляющими разборы анализатора в критических для разметки корпуса точках, например при разметке частотных новых слов[9];

При разметке корпуса со снятой омонимией тексты последовательно обрабатываются:

- автоматически: аналогично предыдущему случаю, связкой автоматического анализатора и фильтров[10];

[7] Приводится вариант xml-представления разметки для корпуса со снятой омонимией, который используется для хранения и обработки текстов корпуса оффлайн. При онлайн-поиске информация о грамматических разборах хранится в виде индексов.

[8] Используется программа «Mystem» (Segalovich 2003; https://tech.yandex.ru/mystem/); релиз для Национального корпуса русского языка выполнен компанией «Яндекс».

[9] Фильтры разработаны А. Е. Поляковым и Д. В. Сичинавой. С их помощью могут добавляться новые или удаляться ошибочные или не встречающиеся в корпусе «паразитические» разборы, ср. разбор формы *какая* как деепричастия.

[10] На первых этапах создания НКРЯ использовался вариант программы «Диалинг» (Сокирко 2004; http://www.aot.ru), который частично прогнозировал правильные разборы омонимичных словоформ; впоследствии от этой опции решено было отказаться, так как ошибки программы трудно было проконтролировать. В 2012—2013 гг. для предварительь-

- вручную: разметчики разрешают морфологическую омонимию во всех оставшихся случаях и просматривают весь текст целиком, исправляя допущенные программами ошибки.

Единообразное представление информации, полученной в результате работы программ и разметчиков, обеспечивает морфологический стандарт, разработанный в 2001—2004 гг. В. А. Плунгяном, Д. В. Сичинавой, Г. И. Кустовой, А. Е. Поляковым и автором этой книги. Стандарт служит теоретической и методологической основой морфологической разметки и включает решения, касающиеся инвентаря морфологических признаков, состава парадигмы лексемы, ее исходной формы, представлений о грамматической норме (какие словоформы считаются стандартными для данной лексемы, а какие аномальными, ср. формы императива *выйди* и *выдь*), приемов идентификации морфологических разборов и проверки правильности разрешения морфологической омонимии.

Разработчики стандарта морфологической разметки исходили из ряда принципов. Во-первых, как уже было сказано, грамматические признаки, приписываемые словоформе, должны быть понятны максимально широкому кругу пользователей и согласоваться с традицией описаний грамматики русского языка. В тех случаях, когда языковое явление допускает несколько трактовок в русле русской грамматической традиции (так называемые «спорные вопросы» русистики: сколько родительных падежей в русском языке — один или два; входит ли форма превосходной степени в парадигму прилагательного; является ли предикатив особой частью речи и т. д.), морфологический стандарт обеспечивает единообразное решение этой проблемы во всем корпусе, причем по возможности такое, которое было бы приемлемо с точки зрения сторонников любой из существующих трактовок.

Во-вторых, всем словоформам корпуса, признанным формами русского языка (а не включенными в русский текст словоформами иностранных языков), должна быть обязательно приписана некоторая грамматическая характеристика. С этим связана большая исследовательская работа разработчиков корпуса по выявлению словоформ, не описываемых нормами русской грамматики и определению их места в составе или вне состава парадигмы слова.

В-третьих, корпус стремится максимально облегчить для пользователя задачи поиска морфологической и лексической информации. Именно этим подходом продиктовано решение, согласно которому потенциальные pluralia tantum типа *взаимоотношения — взаимоотношение* получают две исходных формы.

Четвертый принцип звучит следующим образом: «Не важно, как названо некоторое грамматическое явление, важно, чтобы оно могло быть сформулировано в виде запроса к корпусу». Так, иногда в грамматической традиции существует несколько обозначений для одного и того же грамматического признака, например будущее время (совершенного вида) = непрошедшее время (совершенного вида).

ной автоматической разметки текстов стала использоваться программа «Mystem», адаптация Т. А. Архангельского.

В корпусе в данном случае ярлыком грамматического признака было выбрано «будущее время» как более традиционное. В то же время разработчики понимали, что исследователь русского языка, использующий термин «непрошедшее время», сможет найти все интересующие его употребления, задав два запроса:

наст. время, несов. вид
буд. время, сов. вид[11].

С этих же позиций при выработке решений, касающихся других спорных вопросов грамматики, выбор делался в пользу более дробного представления грамматической категории. Например, в состав парадигмы существительного был включен второй родительный падеж (ср. *спору нет*) с учетом того, что исследователь, считающий это употребление формой дательного падежа, сможет задать запрос:

существительное + второй род. падеж.

Обратное неверно; перечисление всех позиций, в которых встречаются формы «дательного падежа в функции родительного»:

мало / *много* / *недостаточно* / *побольше* / *полкило* / *две тарелки*…
дать / *налить* / *насыпать* / *пожалеть* / *купить* / *попробовать*…
нет / *не хватает* / *не нужно* / *обойтись без* / *осталось* / *жалко*…
наделать / *натерпеться* / *наесться* / *натаскать* / *наговорить*…
+ сущ.: неодуш., м. р, дат. пад.,

создало бы много неудобств пользователю и дало бы некоторое количество «шума», ср. *Предложил коллективу искупаться*.

Пятый принцип можно было бы назвать «Не решай за исследователя». Если контекст не позволяет во фразе *Я тебя буду звать Квазимодо* однозначно определить падеж существительного (именительный *vs*. творительный), то в корпусе сохраняются два альтернативных разбора[12] — в противном случае разметчик корпуса выступил бы в роли, которую надлежало взять на себя лингвисту-исследователю.

Наконец, ряд компромиссных решений был принят, исходя из особенностей технического представления грамматической информации и возможности идентификации грамматических разборов в процессе автоматической разметки. Большинство этих решений касаются аналитических грамматических форм, см. с. 27. Техническими трудностями автоматического определения грамматической информации вызвано соглашение об упрощенном формате разметки корпуса с неснятой омонимией: в нем, частности, отсутствует информация о переходности / непереходности глагола, о форме второго винительного падежа (см. с. 27), помета

[11] Здесь и далее для удобства читателей приводятся русские обозначения морфологических признаков.

[12] В корпусе со снятой лексико-грамматической омонимией.

«инициалы»; помета «сокращение» приписана только наиболее частотным единицам типа «т. п.», «п/п»[13].

Конкретные решения, принятые в морфологической разметке, опираются, прежде всего, на работы (Зализняк 1977/2003; 1967). Далее мы обсудим отступления от модели «Грамматического словаря», продиктованные изложенными выше соображениями.

Трактовка аналитических форм

В корпусе используется в основном пословный принцип морфологической разметки; кроме того, в процессе разработки находится «второй слой» разметки на уровне неоднословных устойчивых оборотов (*в течение*, *во что бы то ни стало* и т. п.; ср. также опыт корпуса ХАНКО (Копотев 2004; Копотев, Мустайоки 2003)). Предусмотрен поиск лексических единиц как в составе оборотов, так и вне их. Например, пользователь, ищущий сочетания предлога *в* с существительным в винительном падеже, выбрав опцию «искать вне оборота», будет избавлен от многочисленных примеров употребления этого предлога в составе сложных предлогов (типа *в течение*) и других оборотов.

Тем не менее аналитические грамматические формы: будущее время несовершенного вида (*будет оценивать*), условное наклонение (*оценили бы*), прошедшее время совершенного вида пассивного залога (*был оценен*), аналитические формы сравнительной степени прилагательных и наречий (*более экзотически*) и нек. др. — разбираются в настоящее время только пословно, т. е. пользователь должен задавать их в поиске как конструкцию из двух элементов.

Так, формы сложного будущего времени кодируются как

быть: буд. время + <глагол>: инфинитив, несов. вид (*буду петь*),

формы условного наклонения — как

<глагол>: прош. время / инфинитив + *бы* / *б* / *чтобы* / *чтоб*,

аналитические формы сравнительной и превосходной степени прилагательных и наречий — с помощью формул

более / *менее* + <прил.>: положит. форма / <наречие>

или

самый / *наиболее* / *наименее* + <прил.>: положит. форма / <наречие>.

---

[13] Заметим, что формам типа *Рис.* помета «сокращение» (ср. *рисунок*) в неснятом корпусе не приписывается, дабы избежать паразитических омонимичных разборов у несокращенных написаний (ср. *рис* как название еды).

«Морфологический» принцип хорош своей относительной простотой и последовательностью: его легко провести программными средствами (для идентификации грамматической формы не требуется обращаться к ее контексту), а предложения, содержащие аналитические формы, вообще говоря, можно найти с помощью стандартных поисковых запросов. Кроме того, это решение уравнивает конструкции типа *будет плакать* с другими близкими инфинитивными конструкциями со значением будущего времени: *станет плакать*, *начнет плакать*, а признанные аналитические формы суперлатива — с похожими, но менее стандартными конструкциями типа *в наибольшей степени заинтересованный* или *менее всех заметный*. Пословный подход также избавляет нас от проблемы, как трактовать расстояние между словами в поиске (например, как задать запрос, если пользователь хочет найти паттерны типа *будет посылать им*, *им будет посылать* и *будет им посылать*).

Как слабую сторону данного решения мы можем отметить наличие «шума» при поиске и расхождение с традицией грамматического описания русского языка. Неудобство при поиске возникает, во-первых, если пользователь, например, ищет формы инфинитива (или прошедшего времени глагола), но не имеет возможности автоматически отсеять аналитические формы. Во-вторых, при поиске самих аналитических форм пользователь должен задавать произвольное расстояние между составляющими из-за свободного порядка элементов конструкции и отсюда велика вероятность получить в выдаче примеры, где искомые формы встречаются случайным образом (ср. *Самым ценным качеством* ***будет*** *именно умение* ***предвидеть***; подробный разбор этих случаев см. в Копотев 2004).

Безусловно, больше всего мы отходим от грамматической традиции в случае форм будущего времени и условного наклонения. Возможный выход мы видим в том, чтобы в будущем разбирать аналитические грамматические формы как особый вид оборотов[14]. От стандартных оборотов они будут отличаться большей свободой лексического наполнения и нежестким порядком входящих в них элементов.

Техническую сложность, кроме того, представляет разметка употреблений сложного будущего времени с однородными формами типа *буду читать*, *писать* (Там же), так называемых сериальных глагольных конструкций (Вайс 1993) типа ***буду сидеть смотреть****, как ты занимаешься*, а также аннотация оборотов типа *должен буду думать*, допускающих две интерпретации:

должен + *думать*: буд. время

и

должен: буд. время + *думать*.

[14] Помимо указанных, сюда войдут сложные формы времени и наклонения неглагольных модальных показателей: *должен был*, *должен будет*, *должен был бы*, *сложнее стало* (*получать визы*), а также предикативов: *ему было безразлично* (*что будет с Ниной*). Интересно, что, например, в корпусе ХАНКО этот подкласс аналитических форм в настоящее время не учитывается.