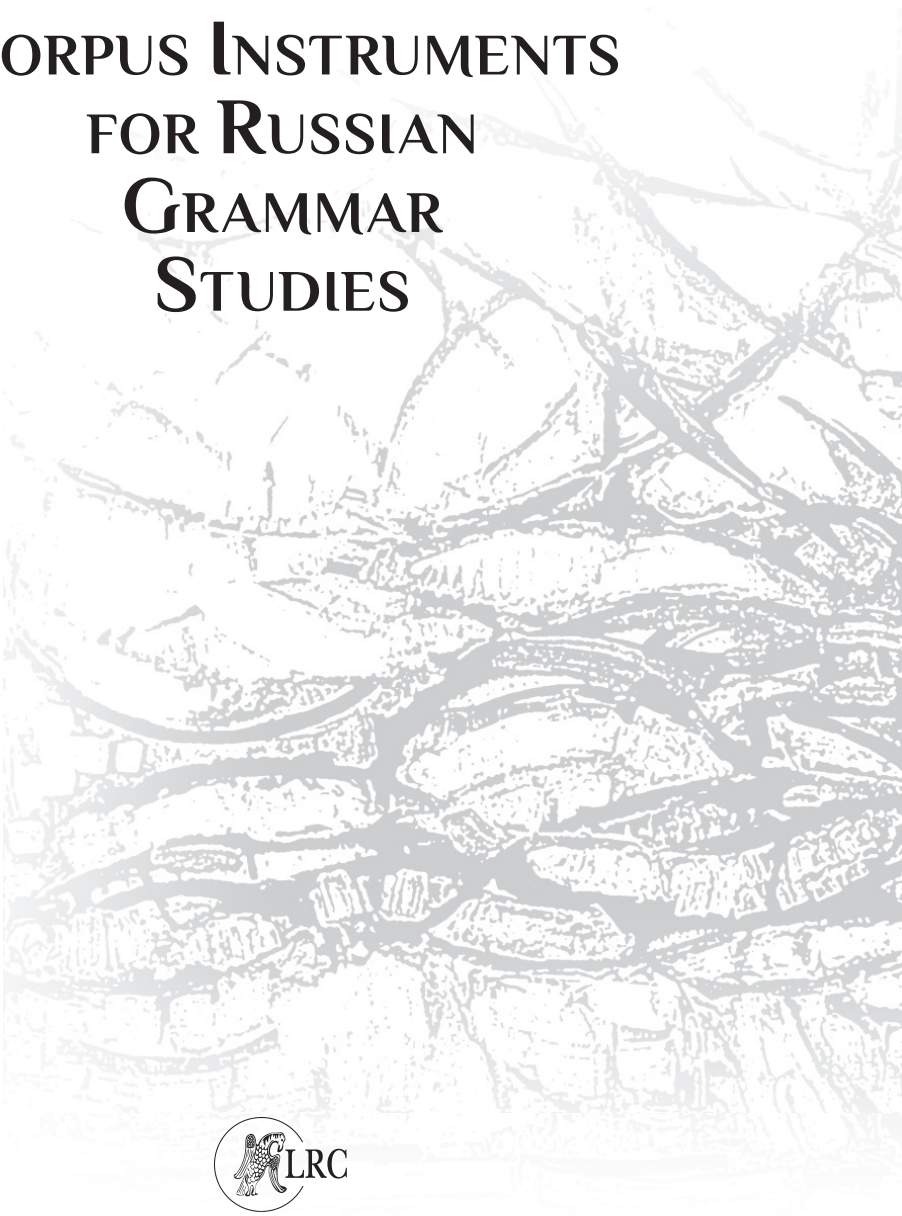
КОРПУСНЫЕ ИНСТРУМЕНТЫ В ГРАММАТИЧЕСКИХ ИССЛЕДОВАНИЯХ РУССКОГО ЯЗЫКА

Olga Lyashevskaya

Corpus Instruments for Russian Grammar Studies



LRC PUBLISHING HOUSE
THE MANUSCRIPT HERITAGE OF OLD RUS
MOSCOW 2016

О. Н. Ляшевская

Корпусные инструменты в грамматических исследованиях русского языка

ИЗДАТЕЛЬСКИЙ ДОМ ЯСК
РУКОПИСНЫЕ ПАМЯТНИКИ ДРЕВНЕЙ РУСИ
МОСКВА 2016

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

Предисловие 7

Часть 1. Развитие корпусных инструментов и технологий

1.1. Национальный корпус русского языка и его аннотация 13

1.2. Словоизменение 19
1.2.1. Морфологический стандарт корпуса 19
1.2.2. Пополнение грамматического словаря по корпусным данным 40
1.2.3. Соревнования морфологических анализаторов 49

1.3. Лексико-семантические классы 64
1.3.1. Принципы лексико-семантической разметки 64
1.3.2. Разрешение лексико-семантической неоднозначности с помощью векторов контекстных маркеров 88

1.4. Интерфейс морфосинтаксиса и семантики 112
1.4.1. Аннотация лексических конструкций в системе ФреймБанк 112
Приложение 164
1.4.2. Распознавание семантических ролей на основе ФреймБанка 176
1.4.3. Автоматическая синтаксическая аннотация корпуса и соревнования парсеров зависимостей 193

1.5. Словообразование 210

1.6. Частотные словари на базе корпуса 224
1.6.1. Частотный словарь современного русского языка 225
1.6.2. Частотный лексико-грамматический словарь 246

Часть 2. Квантитативные подходы к исследованию на корпусных данных

2.1. Векторное представление корпусных данных и профили контекстного «поведения» языковых единиц 257

2.2. Грамматические профили 279
2.2.1. Грамматическая специализация глаголов в формах времени и наклонения 279

2.2.2. К описанию дистрибуции форм единственного и множественного числа имен существительных 319

2.3. Конструкционные профили 338

2.3.1. Конструкционные профили приставочных видовых пар 338

2.3.2. Инкорпорация и экскорпорация в глагольном управлении: участник «часть тела» 358

2.3.3. Инструментальная и генитивная конструкция формы имен существительных 373

2.4. Семантические профили: классы глаголов и выбор видовых приставок. 382

2.5. Радиальный профиль значения: пространственная конструкция с предлогом *поверх* 407

Заключение 430

Приложения

Приложение 1 435
Приложение 2 457
Приложение 3 468
Приложение 4 474

Библиография 480

Принятые сокращения 514
Abstract 517

Предисловие

Корпусная лингвистика — довольно молодое направление лингвистической науки. Национальному корпусу русского языка исполнилось 10 лет, а самому старому представительному корпусу объемом более 100 миллионов словоупотреблений, Британскому Национальному, — всего 25 лет. Прежде всего уточним, что термин «корпусная лингвистика» предполагает два понимания: это и наука о том, как создавать лингвистические корпуса, и методы исследования языка с привлечением корпусных данных. Обычно считается, что созданием корпусов занимаются инженеры и программисты, а исследованиями на данных корпуса — собственно лингвисты. В случае Национального корпуса русского языка это не так: корпус создавался лингвистами и для лингвистов (хотя и с помощью «инженеров»). Мне повезло несколько раз: в начале двухтысячных оказаться в отделе Лингвистических исследований ВИНИТИ РАН, когда только появилась и начала реализовываться идея Национального корпуса; затем в отделе корпусной лингвистики и лингвистической поэтики Института русского языка им. В. В. Виноградова, где ведется основная работа над корпусом; после этого в Институте лингвистики Университета Тромсё, где были начаты первые исследования Национального корпуса с помощью квантитативных методов; и наконец в НИУ «Высшая школа экономики», где собралась замечательная команда исследователей русского языка. Так и получилось, что я работаю в обоих направлениях корпусной лингвистики.

Соответственно, книга, которую вы держите перед собой, тоже имеет две части. Первая часть посвящена лингвистической аннотации текстов Национального корпуса русского языка (ruscorpora.ru) на разных уровнях: словоизменения, словообразования, синтаксиса и семантико-синтаксического интерфейса, лексико-семантических классов. Мы обсуждаем исходные теоретические установки, связанные с системой аннотации, разработку вспомогательных лингвистических ресурсов (словарей и баз данных), компьютерных инструментов разметки и самое интересное — то, что я бы назвала «сопротивлением материала», — описание сложных случаев языкового материала, которые могут вызвать трудности как при автоматической аннотации, так и при ручной разметке. Чуть выходя за рамки задач непосредственно Национального корпуса, мы обращаемся к вопросам стандарта оценки автоматической разметки текстов и рассказываем о двух инициативах в области компьютерной лингвистики — о соревнованиях морфологических и синтаксических парсеров. В конце первой части описываются производные корпуса — частотные словари, которые можно построить на корпусных данных.

Во вторую часть входят работы по исследованию грамматики и лексики русского языка квантитативными корпусными методами. Понятие грамматического «поведения» языковых единиц в применении к корпусу видится как распределение разного рода элементов в контексте. Это грамматический профиль (распределение форм словоизменения), конструкционный профиль (распределение конструкций некоторой «целевой» лексемы), лексический, лексико-семантический профиль (распределение лексем или лексико-семантических классов в контексте другой лексемы или конструкции), радиальный профиль значения (распределение значений / частных употреблений языковой единицы). С помощью методов грамматического, конструкционного, семантического профилирования мы анализируем грамматическую специализацию русских глаголов по формам вида, времени и наклонения; вариативность образования приставочных видовых пар с разными приставками; ограничения на заполнение слотов и связанные с этим вариации значения в генитивной конструкции формы и в пространственной конструкции с предлогом *поверх*. Квантитативные методы, привлекаемые для анализа, разнообразны: от чисто описательных частот и процентных долей до теста Фишера и регрессии.

Создание корпусов и квантитативные исследования, требующие масштабной доразметки корпусных данных, — дело чрезвычайно трудоемкое, и его приятнее делать в коллективе. Поэтому в этом предисловии я бы хотела поблагодарить моих соавторов, с которыми мне посчастливилось работать в наших многочисленных корпусных проектах: В. А. Плунгяна и Д. В. Сичинаву (морфологическая разметка корпуса, см. Ляшевская и др. 2005в) пополнение грамматического словаря, см. (Ляшевская и др. 2007), Е. В. Рахилину, Г. И. Кустову, Е. В. Падучеву, О. Ю. Шеманаеву, Б. П. Кобрицова, Т. И. Резникову (лексико-семантическая разметка корпуса и разрешение неоднозначности, см. Kustova et. al. 2009; Шеманаева и др. 2007; Рахилина и др. 2006), С. Ю. Толдову (синтаксическая разметка корпуса), Ю. Л. Кузнецову, М. С. Кудинова и Е. В. Кашкина (проект ФреймБанк, см. Кузнецова, Ляшевская 2009; Кашкин, Ляшевская 2013; Lyashevskaya, Kashkin 2014), Е. А. Гришину, М. Г. Тагабилеву, И. Б. Иткина, Е. К. Павлову (словообразовательная разметка корпуса, см. Гришина и др. 2009), А. А. Бонч-Осмоловскую, Е. Г. Соколову, С. О. Савчук, С. А. Коваля, еще раз С. Ю. Толдову и команду студентов МГУ (И. Астафьева, А. Королева. М. Ионов, М. Кудринский, Д. Привознов, Евг. Сидорова и мн. др.), с которыми мы организовывали соревнования парсеров (см. Ляшевская и др. 2010; Толдова и др. 2012; Gareyshina et al. 2012; Bonch-Osmolovskaya et al. 2013), С. А. Шарова, моего соавтора по частотному словарю (Ляшевская, Шаров 2009). Вместе с А. В. Десятовой и А. А. Маховой мы делали проект по топологической классификации лексики и исследованию пространственных конструкций (см. Махова и др. 2009; Десятова и др. 2008), с О. А. Митрофановой, П. В. Паничевой, С. В. Романовым, Н. С. Кузнецовой, М. А. Грачковой, А. С. Шиморина и А. С. Шурыгиной — проект по автоматическому разрешению лексико-семантической омонимии, а с В. Г. Сибирцевой и Н. В. Карповым — проекты по использованию материалов корпуса в учебных целях. Наконец, самые большие слова благодарности — основа-

телям исследовательской лаборатории CLEAR group Университета Тромсё Л. Янде, Т. Нессету, С. В. Соколовой, (снова) Ю. Л. Кузнецовой, А. Б. Макаровой и А. В. Эндресен (Байдимировой), вместе с которыми мы учились применять квантитативные корпусные инструменты к данным Национального корпуса русского языка. Я еще раз благодарю своих соавторов за любезное разрешение использовать материалы наших совместных статей в этой книге. Первоначальные варианты многих глав были опубликованы в материалах конференции «Диалог» — и мы бесконечно благодарны ее организаторам и слушателям за многолетний интерес к публикациям разработчиков Национального корпуса.

Особенные слова должны быть посвящены светлой памяти безвременно ушедшего И. В. Сегаловича. Илья одним из первых поддержал идею Национального корпуса, щедро делясь своей позитивной энергией и креативными идеями на семинарах разработчиков корпуса. По инициативе Ильи «Яндекс» стал основным техническим партнером корпуса и инициировал исследовательские гранты, с помощью которых были проведены первые математические исследования на материалах корпуса. Тут же мы должны произнести много теплых слов благодарности в адрес других сотрудников компании «Яндекс», которые на протяжении более десятка лет обеспечивают техническую поддержку корпуса и терпят все капризы лингвистов-разработчиков: А. И. Зобнина, И. Е. Шалыминова, Н. В. Григорьева, А. В. Сокирко, А. А. Аброскина, В. А. Титова, С. А. Григорьеву, Е. С. Грунтову и др. И еще: огромное спасибо студентам трех московских вузов, МГУ, РГГУ и НИУ ВШЭ, принимавших участие в наших проектах в качестве разметчиков. Корпус не был бы таким, какой он есть, без ваших усилий.

В европейской традиции принято благодарить не только научных руководителей, начальников, учителей и коллег, но и тех, с кем просто пил чай. Я бы хотела поддержать эту прекрасную традицию и назвать тех, кто был рядом, помогал, спасал, создавал хорошее творческое настроение и беседовал за чаем о лингвистике и не только: Ю. Родина, М. Пост, Д. Пинеда, П. Иосад, М. Панчева, Х. Андреассен, Л. Антонсен, Р. Михайлык, М. Нордрум, Д. Папрот, Т. Горностай, А. Недолужко, А. Бердичевский, Х. Экхофф, А. Рубин, О. Урюпина, М. Кронгауз, М. Даниэль, Н. Добрушина, Е. Добрушина, В. Апресян, Б. Орехов, Т. Архангельский, Ю. Ландер, А. Летучий, Я. Ахапкина, Д. Алексеевский, О. Виноградова, А. Марушкина, Т. Никитина, Н. Слюсарь, В. Файер, Ю. Галямина, Ю. Кувшинская, М. Худякова, Т. Ряпина, Н. Зевахина, С. Князев, Б. Иомдин, Н. Стойнова, П. Браславский, П. Аркадьев, С. Сай, М. Овсянникова, А. и Л. Ландманы, И. Микулинская, Л. Кацман, В. Гусев, Н. Галицкая, С. Бурлак, В. Степанов, Т. Михайлова, Е. Марголис, Б. Кротов, Е. Калинина, В. Цуканова, Г. Дурново, Н. и А. Горовые, Н. и О. Сидоренковы, Е. Шаульский, А. Занадворова, Е. Ягунова, Л. Пивоварова, М. Копотев, М. и А. Беловы, И. и Ю. Ребриковы, Е. и А. Ребриковы и многие, многие другие. В заключение я хочу произнести слова признательности моим родителям Н. С. и Н. Ф. Ляшевским, моему мужу Саше и сыновьям Егору и Степе. Спасибо вам за терпение, сочувствие и поддержку.

Текст всей книги внимательно прочитали А. Ч. Пиперски, Е. В. Ягунова, А. Я. Шайкевич и официальные рецензенты М. Р. Пентус и И. В. Азарова. Я бесконечно благодарна им за вдумчивые замечания и уточнение ряда формулировок. Безусловно, все оставшиеся несообразности — недоработка автора. Моя глубокая благодарность В. В. Столяровой, Е. Г. Сметанниковой, И. В. Богатыревой, осуществившим техническую подготовку издания к печати.

* * *

Рукопись монографии подготовлена при поддержке Научного фонда НИУ ВШЭ, индивидуальный исследовательский проект № 14-01-0069, 2014-2015. Издание осуществлено с помощью издательского гранта Фонда фундаментальных лингвистических исследований, грант № В-28, 2014/2015 гг.

ЧАСТЬ 1

РАЗВИТИЕ КОРПУСНЫХ ИНСТРУМЕНТОВ И ТЕХНОЛОГИЙ

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,