*О. Н. Ляшевская, В. А. Плунгян, Д. В. Сичинава*

# О МОРФОЛОГИЧЕСКОМ СТАНДАРТЕ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА*

## 0. Введение

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 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

---

* Данная статья представляет собой переработанный и расширенный вариант статьи, опубликованной в журнале «Научно-техническая информация. Сер. 2: Информационные процессы и системы», № 6, 2005, с. 2-9.

тем, что позволяет разметить огромные массивы текстов без участия человека (программа-парсер 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 именительный падеж употребляется после предлога *за* (ср. *Что **за прелесть** эта Наташа*). Таким образом, формально-морфологический подход предлагает совершенно иной взгляд на грамматику русского языка, идущий вразрез со сложившейся грамматической традицией, и это противоречие делает такой корпус малопригодным, в частности, 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 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 лингвистическом коллективе, разрабатывающем Национальный корпус русского языка, 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; тем самым, морфологическая разметка больших по размеру корпусов может быть осуществлена, по крайней мере в значительной части, при помощи компьютерной программы.

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.) 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 челове-

[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.

ка[2]. 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 морфологических характеристик, 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.

Представим кратко технологию морфологической разметки, применяемую в Корпусе.

### 1. 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:

1) русские словоформы (в том числе неопознанные и гипотетические словоформы) — состоят из букв кириллицы и, в редком случае, из знаков дефиса (-) и апострофа (’): *человек, что-то, д’Артаньян*;

2) 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*;

3) иноязычные фрагменты текста — состоят из словоформ, записанных латинскими, греческими и другими некирилличе-скими буквами (How do you do, π), или из кириллических сло-

[2] Пользуясь случаем, выражаем благодарность Д. С. Ганенкову, Е. А. Гришиной, О. В. Драгой, С. А. Ковалю, Г. И. Кустовой, Ю. А. Ландеру, Т. А. Майсаку, Н. В. Перцову, А. Е. Полякову, Ю. Д. Семьяновой, С. В. Уляхиной и др., принимавшим участие в морфологической разметке корпуса или экспертизе ее результатов.

воформ, представляющих запись текста на иностранном языке (*Гуд ивнинг, Здоровеньки булы*)[3];

4) знаки препинания: точка, запятая, тире, кавычки, вопросительный, восклицательный знак, двоеточие, многоточие и нек. др.;

5) прочие символы типа %, >, $ и др.;

6) команды мета-разметки и структурной разметки текста в угловых скобках, например:

<meta name=«author» content=«Гроссман Василий»>

,  (начало и конец абзаца).

Все фрагменты текста, кроме русских словоформ, а в корпусе со снятой грамматической омонимией — ещё 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. Деривационные признаки включены в состав семантической разметки, представляющей собой расширение морфологической аннотации[4]. 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 «полей», или групп помет:

---

[3] Словоформы, записанные смесью кириллических, латинских и прочих символов (*e-mail'ы, PRumь*), в настоящее время условно считаются относящимися к иноязычным, хотя кириллические элементы в их написании говорят как раз в пользу адаптации недавних заимствований к грамматической системе русского языка и о появлении у них словоизменения. В связи с этим в ближайших планах развития морфологической разметки Корпуса предусматривается разработка специальных средств аннотации словоформ такого типа.

[4] См. статью Г. И. Кустовой, О. Н. Ляшевской, Е. В. Падучевой, Е. В. Рахилиной «Семантическая разметка лексики в национальном корпусе русского языка: принципы, проблемы, перспективы» в настоящем сборнике.

1. Лексема, которой принадлежит словоформа (указывается «словарная запись» данной лексемы);

2. Множество грамматических признаков данной лексемы, или словоклассифицирующие характеристики (указывается принадлежность лексемы к той или иной части речи, а также, например, род для существительного, переходность для глагола и т. п.)[5];

3. Множество грамматических признаков данной словоформы, или словоизменительные характеристики (например, падеж для существительного, число для глагола);

4. Информация о нестандартности грамматической формы, орфографических искажениях, сокращенном написании типа *млн*, *г-н* и т. п.

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 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!*:

```
<w lex='вы' gr='S-PRO,pl,2p=nom'>Вы</w>
<w lex='оставить' gr='V,pf,tran=act,fut,2p,pl=distort'> оста-авите</w>!
```

[*Александр Солженицын. В круге первом* (*т. I*)]

Как уже было сказано, тексты Корпуса размечаются автоматически (по крайней мере, на первом этапе) с помощью специальных программ-парсеров. При разметке используются встроенные в эти

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

[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*.

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 омонимией осуществляется с помощью программы «Mystem» (автор И. В. Сегалович, компания «Яндекс»), порождающей все возможные разборы словоформ, а также гипотезы относительно словоформ, отсутствующих в словаре. При разметке корпуса со снятой омонимией тексты последовательно обрабатываются:

1) программой «Диалинг» (коллектив авторов под руководством А. В. Сокирко, группа «Диалинг», www.aot.ru), которая частично прогнозирует правильные разборы омонимичных словоформ;

2) автоматическими фильтрами (авторы А. Е. Поляков и Д. В. Сичинава), указывающими правильные разборы для нераспознанных ранее случаев омонимии в самых частотных контекстах;

3) 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. авторами настоящей статьи (с участием Г. И. Кустовой и А. Е. Полякова). 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 *пойди* и *поди*), 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 как можно большему широкому кругу пользователей и согласовываться с традицией описаний грамматики русского языка. В тех случаях, когда языковое явление допускает несколько трак-

товок 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, которые 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 вид и залог глагола считаются двойственными категориями: словоклассифицирующими и словоизменительными. Так, например, словоформа *открылся* входит, с одной стороны, в парадигму лексемы *открыться*, а с другой, — в расширенные парадигмы глаголов *открыть* — как форма среднего (медиального) залога и *открывать*(*ся*) — как форма совершенного вида. Лингвист, изучающий семантику глагола, получит при поиске по заглавному слову *открыться* также и формы от *открыть* и *открывать*(*ся*); исследователь же глагольного вида или залога, выбрав соответствующий параметр, может ограничить свой поиск нужным элементом видовой или залоговой пары.

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). 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 вре-

мя», сможет найти все интересующие его употребления, задав запрос:

наст. время, несов. вид + буд. время, сов. вид.

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:

дат. падеж + 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 буду звать тебя Квазимодо* однозначно определить падеж существительного (именительный *vs.* творительный), то в Корпусе сохраняются два альтернативных разбора — 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. раздел 2. 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. раздел 4), пометы «инициалы» и «сокращение».

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

## 2. Трактовка аналитических форм

В Корпусе в настоящее время 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. Как особый вид оборотов в будущем предполагается также разбирать 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.

На уровне пословной разметки аналитические формы получают «морфологическую» трактовку. 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,*

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>[7].

[7] Здесь и далее в статье для удобства читателей приводятся русские обозначения морфологических признаков.

«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 другой близкой инфинитивной конструкцией со значением будущего времени: *станет плакать*, а признанные аналитические формы суперлатива — с похожими, но менее стандартными конструкциями типа *в наибольшей степени заинтересованный* или *менее всех заметный*.

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. Как уже было замечено, выход мы видим в том, чтобы разбирать аналитические грамматические формы как особый вид оборотов[8]. 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* [Копотев 2004], так называемые сериальные глагольные конструкции [Miller 1970; Вайс 1993] типа

[8] 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.

***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*

На двух уровнях, пословном и на уровне оборотов, предполагается разбирать также разрывные формы отрицательных и неопределенных местоимений типа *ни у кого*, *кое с кем*, взаимные местоимения типа *друг с другом*, составные числительные типа *триста двадцать пять* и аналитические формы императива типа *давайте споём*.

### 3. Части речи

Морфологический стандарт Корпуса включает 16 частеречных характеристик: имена существительные, прилагательные, числительные, числительные-прилагательные, глаголы, наречия, предикативы (*вам **пора** ужинать*), вводные слова, местоимения-существительные, местоимения-прилагательные, местоимения-предикативы (***нечего** тебе там делать*), местоимения-наречия, предлоги, союзы, частицы, междометия. Список частей речи в целом совпадает с используемым в «Грамматическом словаре», за исключением категории «местоименное наречие» (*там*, *сколько-нибудь*, *по-вашему*). Напомним, что подход, принятый в Грамматическом словаре, представлял собой известный компромисс. А. А. Зализняк пишет: «Все прочие слова, кроме имен и глаголов, образуют один грамматический разряд, где парадигма состоит из единственной формы... Вопрос о разделении этих слов на части речи, как известно, весьма сложен. Поскольку, однако, для словоизменения это несущественно, в настоящем словаре не предлагается какого-либо самостоятельного решения данного вопроса, а используется практически та же система рубрик, что в современных толковых словарях... Это разделение носит в сущности синтаксический характер» [Зализняк 1977/2003, с.8].

Включение в номенклатуру частей речи Корпуса категории «местоименное наречие» по семантическим и отчасти морфологическим критериям (местоименные наречия относятся к разряду наречий, не имеющих форм сравнительной степени) является дальнейшим сближением с лексикографической традицией (ср., например, [Ожегов, Шведова 1999; Кузнецов 2002 и др.]).

С другой стороны, наречия «Грамматического словаря», полностью совпадающие с падежными формами существительных (типа *утром*), в Корпусе, вопреки грамматической традиции, не выделяются (соответствующие единицы разбираются как существительные).

Предлог *ради*, имеющий в «Грамматическом словаре» статус «предлог; послелог», относится в нашем стандарте к категории предлогов. Поиск употребления *ради* в функции послелога (в контекстах типа *справедливости ради*) можно задать с помощью простого запроса:

сущ. в род. падеже + *ради*; расстояние между словами: 1.

Единая трактовка словоформы *ради* как предлога позволяет также не навязывать своего решения исследователям в таких спорных случаях, как *нашего ради спасения*:

(*нашего ради* [посл.]) *спасения* vs. *нашего* (*ради* [предл.] спасения).

### 4. Падежная система

Помимо шести основных падежей [Русская грамматика 1980], в Корпусе выделяются звательный, второй родительный, второй предложный, второй винительный падежи, а также счетная форма[9].

Значения второго родительного и второго предложного падежей приписываются всем существительным мужского и женского рода (типа *мёд, жир, даль* и др.), у которых отмечена соответствующая особенность парадигмы. Помета о наличии второго предложного тем более необходима, что для многих слов она кодирует форму, отличающуюся от формы дательного падежа только ударением на окончании (ср. *к мёду* и *в меду́*, *поклониться те́ни* и *в тени́*), что немаловажно для адекватной работы акцентуационного модуля в Корпусе.

Значение второго винительного падежа[10], полностью совпадающего у одушевленных существительных и числительных по форме с именительным падежом (ср. *идти в* ***солдаты***, *по* ***два*** *мальчика*, *ходить по* ***двое***)[11], приписывается в корпусе со снятой омонимией вручную разметчиком, просматривающим все «подозрительные»

[9] См. [Зализняк 1967, с. 43-52].

[10] Ср. также термин И. А. Мельчука «винительный с потерей одушевленности» [Мельчук 1995].

[11] См. [Зализняк 1967, с. 50-52; 13].

случаи употребления именительного падежа после предлога. Данное техническое решение позволяет, с одной стороны, отделить такие необычные случаи от других употреблений номинатива, а с другой, — избежать избыточной омонимии в формах именительного падежа в корпусе с неснятой омонимией (доля употреблений второго винительного падежа пренебрежимо мала по сравнению с частотностью форм собственно номинатива).

Проблема идентификации формы, идентичной форме именительного падежа единственного числа, возникает и при разметке сложных числительных типа *в **одна тысяча девятьсот** сорок пятом году*. Однако эта форма сохраняется после любых предлогов и при любом падеже последней (склоняемой) составляющей числа, ср. *с девятьсот пятнадцатого года*, поэтому в корпусе было принято решение придать ей особый статус — несогласуемой формы без падежного показателя, ср.:

| | | |
|---|---|---|
| одна = A-NUM = f,sg, | ср. *одна вещь* | одна = A-NUM = f,sg,nom; |
| тысяча = S,f,inan = sg, | ср. *до тысячи* | тысяча = S,f,inan = sg,gen; |
| девятьсот = NUM, | ср. *в девятистах* | девятьсот = NUM = loc[12]. |

Формами звательного падежа считаются словоформы, употребленные в функции обращения и отличающиеся по внешнему виду от форм номинатива. К ним относятся как реликты древнерусского вокатива (очень частотные *Боже* и *Господи* и единичные формы других слов типа *старче*, *друже, княже* и т. п.), так и новые формы с усечением флексии *-а* (типа *Мить*, *Зойк, мам, ребят*). Формы с растянутым корневым гласным типа *Ми-и-итя*, сохраняющие флексию номинатива, считаются формами именительного падежа с «орфографическим искажением». Поиск таких форм возможен с помощью запроса:

сущ: им. пад + distort.

Проблема «счетной формы» для словоформ *часá, шагá, рядá, шарá* [Зализняк 1967, с. 46-48] появилась в Корпусе с внедрением акцентуационной разметки: за исключением места ударения эти формы совпадают с формами родительного падежа единственного числа. Признак «счетной формы» добавляется к разбору «род. пад. ед. ч.» (с сохранением последнего) в корпусе со снятой омонимией в сочетаниях указанных лексем с числительными *два, три, четыре,*

[12] В устной речи встречается также беспадежная форма существительного *ноль* (например, при произнесении дат): *пятого ноль первого девяносто шестого* (5.01.96).

*полтора, пол*[13], ср. равно возможные варианта *два ря́да* и *два ряда́*; только у слова *час* флексионное ударение считается единственно возможным (но акцентная вариативность признается у этого слова в сочетании *четверть часа*).

Следует заметить, что счетные формы представляют собой одну из реализаций более общего морфологического явления — обязательного или факультативного сдвига ударения на окончание, который свойствен значительному числу словоформ второго предложного падежа (*в пыли́*), а у лексемы *шар* — в творительном падеже в выражении *хоть шаро́м покати*. Для слова *час* признак счетной формы, как было сказано выше, факультативно приписывается также в выражении *четверть часа́*.

Таким образом, счетная форма встраивается в систему реляционных падежей, и ее можно считать «третьим» родительным падежом:

| ***Падеж*** | ***Совпадение с другим основным падежом*** | ***Сдвиг ударения*** | ***Функции падежа пересекаются с функциями:*** |
|---|---|---|---|
| второй род. п. | > дат. п. | — | род. п. |
| второй вин. п. | > им. п. | — | вин. п. |
| второй предл. п. | > дат. п. | (+) | предл. п. |
| счетная форма | > род. п. | + | род. п. |

Обязательный сдвиг ударения на окончание наблюдается также у лексем *след, чёрт* и нек. др. в выражениях типа *без следа́*, *ни следа́*, *нет / не осталось / не отыщешь и следа́* [Зализняк 1977/2003], и здесь мы, по-видимому, имеем дело с еще одним гибридным падежом, чье значение вкладывается в значение второго родительного падежа (партитивное употребление), а форма совпадает со счетной формой.

Вместе с тем, в Корпусе признано нецелесообразным выделять признак так наз. «стандартной счетной формы» [Зализняк 1967, с. 288], т. е. употребление форм родительного падежа единственного числа, родительного падежа множественного числа и имени-

[13] *Пол* считается самостоятельной лексемой в соответствии с [Зализняк 1967, с. 78].

тельного падежа множественного числа после названных числительных: *два города*, *две жены*, *две новых булочных / две новые булочные*. Сохранение исходных падежных ярлыков позволяет, в частности, проследить новые тенденции в употреблении форм в этой конструкции [Corbett 1993], ср. примеры из Корпуса[14]:

*За два последние года сюда не заглянула ни одна кинопередвижка* [Александр Яшин. Вологодская свадьба (1962)];

*Мы видим, как три эти блюда постоянно клубятся, дымятся и завихряются в полном беспорядке, и не можем нащупать в них ни смысла, ни логики, ни системы* [Юлия Калинина // «Московский комсомолец», 2003.05.17].

В работе [Еськова 1983] счетной формой признаются еще и формы с нулевой флексией типа *пятнадцать* ***килограмм***, *пять* ***вольт***, заменяющие в количественной конструкции формы родительного падежа множественного числа и совпадающие с формой именительного падежа единственного числа. В Корпусе принято решение считать эти формы аномальными вариантами родительного множественного, если в парадигме лексемы присутствуют также формы родительного множественного на *-ов / -ев* (ср. *пять килограмм = пять килограммов*). У словоформы *вольт* усеченная форма является полноправным членом парадигмы, так как соответствующей формы с окончанием *-ов* не существует. Супплетивная форма *лет* (от лексемы *год*) считается формой родительного падежа множественного числа наряду с формой *годов*, с дополнительным распределением по контекстам (ср. *сорок* ***лет***, но *до сороковых* ***годов***).

С формальной точки зрения, в русском языке, строго говоря, можно было бы постулировать еще один дополнительный падеж («второй дательный») — у числительных *столько, сколько, несколько, много* после предлога *по*: *по стольку, по скольку*, *по нескольку, по многу* (*раз*); ср. стандартную форму дательного падежа *по столь**ким*** и стандартную форму винительного падежа *по столь**ко***. Грамматический словарь признает здесь вариативность форм в счетной конструкции: *по стольку // по столько дней*, однако заметим, что формы *стольку* и *скольку* употребляются также в составе оборота *по стольку по скольку* (впрочем, чаще встречается слитное написание: *постольку поскольку*; ср. также *мало помалу* и нек. др. выражения, в

[14] Из 23 употреблений определений в форме именительного, вместо нормативного родительного, падежа, обнаруженных в Корпусе, 9 принадлежат местоимению *этот* и 5 — прилагательному *последний*.

современном языке орфографически и морфологически трактуемые как наречия; с другой стороны, в текстах встречается слитное написание и в конструкциях типа *помногу часов*). Так как круг лексем, которых касается данное явление, насчитывает всего четыре единицы, а употребление формы на *-у* жестко ограничено контекстом с предлогом *по*, мы всё же предпочли не перегружать грамматическую систему именного склонения новым падежом, а трактовать соответствующие формы как аномальные формы винительного падежа[15].

### 5. Вторая форма повелительного наклонения

В парадигме глаголов в повелительном наклонении различаются формы 2 лица единственного числа, 2 лица множественного числа, и (для глаголов совершенного вида) формы инклюзивного императива (грамматическая помета imper2), совпадающего с формой 1 лица множественного числа будущего времени (*пойдем*). Дополнительная клетка парадигмы выделяется для инклюзивной формы с суффиксом *-те: пойдемте, идемте, споемте, разойдемтесь*. Ее значение находится в привативной оппозиции к значению формы без *-те* (*пойдем, идем, споем, разойдемся*) и обозначает побуждение ***нескольких*** собеседников к совместному действию [Буслаев 1959; Виноградов 1972]. Формант *-те* следует перед возвратным показателем *-ся*, что также говорит в пользу трактовки этой словоформы как словоизменительной[16].

### 6. Форма сравнительной степени на *по-*

В морфологическом стандарте Корпуса, в отличие от большинства описаний русской морфологии (в том числе и от «Грамматического словаря») выделяется как словоизменительная также форма сравнительной степени, отличающаяся от стандартной наличием приставки *по-*: *побольше, поаккуратнее* (*-ей*), *повнимательнее* (*-ей*). Обычно приставка трактуется здесь как элемент, привносящий значение аттенуатива (‘слегка’). В пользу словоизменитель-

[15] Ещё одна морфологически возможная трактовка, к тому же поддерживаемая диахроническими фактами, — анализ этих форм как содержащих показатель дательного падежа *единственного* числа — является проблематичной с семантико-синтаксической точки зрения.

[16] Другие варианты форм наклонения — с флексией *-и* вместо *-ь* и наоборот (в графической реализации) и с аффиксом *-ся* вместо *-сь*: *положь, не боись, избави Боже, садися* — считаются аномальными формами императива.

ной трактовки такой формы говорит полная регулярность её образования, а также то, что приставка не создаёт здесь новой лексемы (**побольшой, *поаккуратный*), что, очевидно, ожидалось бы, если бы морфема *по-* имела словообразовательный статус.

### 7. Отпричастные образования с «не», «полу» и др.

При автоматическом анализе возникает одна своеобразная проблема, связанная скорее с особенностями русской орфографии, чем русской морфологии; тем не менее, на морфологические решения, принимаемые при разметке, это обстоятельство не может не влиять. Речь идет о формах причастий, в качестве первого компонента содержащих либо отрицание *не-* (*неопохмелившийся*), либо адвербиальный компонент типа *полу-* (*полуодетый*), *ново-* (*новоприбывший*), *свеже-* (*свежевыбритый*) и т. п. Слитное написание здесь, так сказать, скрывает лексемную принадлежность этих форм; для того, чтобы форма, например, *неопохмелившийся* опознавалась как принадлежащая лексеме *опохмелиться*, необходимо ввести дополнительное правило разбиения подобных слитных словоформ в письменном тексте: *не* + *опохмелившийся*; аналогично, *свежевыбритый* ⇒ *свеже* + *выбритый*. Таким образом, процедура морфологического анализа строится по образцу разбора других глагольных комплексов, таких, как личная форма глагола с отрицанием или наречием, форма краткого причастия с отрицанием (*не опохмелился, прежде утверждавшийся, не одет*) и др.

Безусловно, это лишь одна из возможных трактовок нетривиального морфологического явления (в частности, можно ставить вопрос о том, нет ли здесь особой разновидности глагольной инкорпорации); мы приняли данное решение, исходя из технической простоты его воплощения в морфологическом анализаторе. Впрочем, правило условного разбиения слитных словоформ может оказаться полезным и для анализа текстов с «плохой орфографией», ср. *нехочу*, *порусски, идуспать* и др. В настоящее время с проблемой таких текстов приходится считаться, поскольку их число постоянно растет (особенно в области современной электронной коммуникации); более того, нарушения орфографических норм в некоторых типах текстов используются и в качестве сознательного стилистического приема, особой языковой игры.

### 8. Вид и залог глагола

Морфологический стандарт Корпуса трактует глагольный вид как категорию, переходную от словоизменительной к словоклассифицирующей (см. раздел 1). Что касается залога, то в Корпусе различаются, по сути, две его подкатегории. Первая из них характеризуется противопоставлением «активный *vs.* пассивный» залог у действительных и страдательных причастий и является словоизменительной. Вторая разновидность залога противопоставляет невозвратные и возвратные глаголы как активные и медиальные и является словоклассифицирующей. Неразличение собственно пассивных употреблений глагола (ср. *Графа заполняется преподавателем*) и декаузативных (ср. *Окно медленно открылось*) обусловлено как техническими трудностями определения семантики словоформ на *-ся*, так и принципом ненавязывания пользователю дискретных решений в спорных случаях.

Как и видовые корреляты, формы противоположного залога, не являясь в точном понимании членами парадигмы глагола, входят в состав «расширенной парадигмы» [Плунгян, Сичинава 2004][17]. Рядом с исходной формой каждого глагола, входящего в состав видовой[18] и/или залоговой пары, приписываются исходные формы связанных глаголов. Например, словоформе *открывался* приписывается исходная форма *открываться*, а также связанные формы *открыться*; *открывать*; *открыть*.

### 9. Pluralia tantum и другие формы множественного числа

Имена pluralia tantum получают разбор, где исходной является форма множественного числа, а помета множественного числа находится среди словоклассифицирущих помет:

| | |
|---|---|
| *часы* | часы = S,m,inan,pl = nom; |
| *из сливок* | сливки = S,inan,pl = gen. |

В то же время у существительных, имеющих формы единственного числа, числовая помета заносится в словоизменительную часть грамматического разбора, ср.:

---

[17] Глаголы, не имеющие форм без *-ся,* признаются глаголами media tantum, с соответствующим признаком в словоклассифицирующей части грамматической пометы.

[18] Согласно [Зализняк 1977/2003], в Корпусе используется «узкое» понимание видовой парности: в число видовых не включаются пары типа *писать—написать* и *сверкать—сверкнуть*.

| | |
|---|---|
| *кислород* | кислород = S,m,inan = nom, sg; |
| *на колесницах* | колесница = S,f,inan = loc, sg. |

Таким образом, подобно глагольному виду, морфологический стандарт Корпуса трактует число как переходную категорию.

В отличие от решения, принятого в «Грамматическом словаре» [Зализняк 1977/2003], формы типа *сапоги* со значением ‘пара предметов’ считаются принадлежащими к парадигме лексемы ед. числа:

| | |
|---|---|
| *сапоги* | сапог = S,m,inan = pl,nom. |

Это связано с тем, что практически любая форма множественного числа существительных допускает интерпретацию как «нерасчленной совокупности» (ср. *паруса* как ‘набор парусов’) или привносит какую-либо иную добавку в значение, выражаемое формой единственного числа (ср. *холод* и *холода*; *решение* и *решения*; *он враг* и *они враги* ‘каждый является врагом другого’). Корпус предоставляет исследователям возможность самостоятельно разобраться в трактовке таких случаев.

Особое решение было принято относительно так называемых «потенциальных» pluralia tantum [Чельцова 1976] типа *раскопки, боеприпасы*. Для ряда слов сама задача указать исходную форму в единственном числе могла бы поставить пользователя в тупик, ср. *тапочек* или *тапочка*; *шпрот* или *шпрота*? Эта проблема решается так же, как и проблема вида — с помощью понятия расширенной парадигмы.

Для плюральных словоформ указывается исходная форма во множественном числе, а также соотносительная форма единственного числа:

| | |
|---|---|
| *сидел без боеприпасов*: | боеприпасы; боеприпас = S,m,pl,inan=gen; |

а для сингулярных словоформ — то же в другом порядке, ср.

| | |
|---|---|
| *захватил с собой боеприпас*: | боеприпас; боеприпасы = S,m,inan=sg,acc. |

Такое решение позволяет избежать потери данных при поиске; в то же время, статус лексем единственного и множественного числа как связанных отличает этот случай от «настоящей» омонимии лексем единственного и множественного числа типа *час* и *часы*.

## 10. Заключение

Мы представили краткий обзор решений, принятых на современном этапе существования Национального корпуса русского языка. Как можно видеть, в некоторых случаях принятие той или

иной грамматической трактовки фактов русского языка было обусловлено скорее техническими причинами, но в большинстве случаев составители Корпуса стремились следовать определенным теоретическим принципам, обеспечивающим информативность и эффективность поиска словоформ и конструкций по заданным грамматическим свойствам — и в то же время не входящих в слишком большое противоречие с существующей традицией.

Следует сказать и еще об одной важной проблеме, возникающей при попытке осуществить полную грамматическую разметку современных русских текстов. Даже если ограничиться современными письменными текстами, представляющими литературный русский язык, наблюдаемая в них степень грамматической вариативности окажется существенно выше той, которую отмечают грамматики русского языка. Помимо того, что в текстах встречаются искаженные написания (об этом говорилось выше), в них также проникают диалектные, региональные, разговорные, жаргонные и т. п. грамматические варианты. И если описательная грамматика русского языка всегда имеет возможность оставить какие-то варианты за пределами рассмотрения (присвоив им ярлык «ненормативных» или вынеся «за ромб» и не дав никакой грамматической характеристики[19]), то корпусная лингвистика работает совсем в другом идеологическом поле: она *обязана* учитывать любые варианты, встреченные в текстах, поскольку они по определению принадлежат корпусу и, тем самым, должны получить адекватный разбор.

По данным Национального корпуса, наиболее регулярно в текстах встречаются следующие отклонения от современной письменной нормы:

1) редукция конечного гласного (ср. *прям, тож, вродь, спасиб*; в этом же ряду можно отметить новые формы вокатива типа *Маш*, которые учитываются в нашей системе морфологической разметки);

2) отпадения конца слова (ср. *оч* вместо *очень*, *лан* вместо *ладно*) и стяжения (ср. *тыща, сёдни, быр(р)о, бушь, всё-тки, кто-*

[19] Ср. практику «Грамматического словаря» Зализняка, в котором «за ромб» выносится информация об «аномальных» формах лексем в составе устойчивых оборотов, например, об употреблении формы *свеч* (вместо *свечей*) в выражении *игра стоит свеч*; аналогично трактуются там и формы типа *по стольку*, о которых шла речь выше.

*нить*; сосуществующие ряды «полных» и очень распространенных «стяженных» форм в склонении личных и вопросительных местоимений типа: *тебя* и *тя*, *тебе* и *те*, *что* и *чё* и др.);

3) морфонологические или морфологические отклонения от стандартных моделей словоизменения (отсутствие переходного смягчения у форм типа *пылесосю*, отсутствие палатализации в формах типа *кудахтая*, контаминация типов склонения в таких формах, как *герлов, сомнамбулов*);

4) вообще вариативность основ, в том числе орфографическая, например, у существительных на *-ние* vs *-нье, -тие* vs *–тье* (*пение* и *пенье*, *счастие* и *счастье*), колебания в написании дефиса у слов типа *квазинаучный* и *квази-научный*, неустойчивая орфография сленговых элементов, не фиксируемых нормативными словарями (например, *галимый* и *голимый*, *флейм* и *флэйм*), и т. п.;

5) широко распространенные нестандартные формы деепричастий на *-а /-я*, *-ась /-ясь* (*положа, наклоня, прислонясь*) наряду с несколько более редкими, но также фиксируемыми старыми вариантными и диалектными формами на *-ши, -чи* (*положивши, выпимши, вышедши, глядючи, сидючи, жалеючи*);

6) склоняемые краткие формы прилагательных в устойчивых оборотах и имитациях фольклорных текстов: *средь бела дня, под белы ручки, на босу ногу, к едрене фене, красну девицу, сладку ягоду*;

7) все большее распространение «неоформленных» имен (т. е. таких, которым не приписывается никакой граммемы падежа), ср. уже отмеченные выше составляющие сложных числительных, а также употребления типа *от Марь Петровны, в святая святых, система исполнитель-заказчик*).

Таким образом, для адекватного описания морфологии текстов Корпуса оказывается необходима модель, учитывающая постоянную и высокую морфологическую вариативность. Парадоксальным образом, подобные модели разрабатываются обычно не применительно к стандартизованным письменным языкам, а применительно к бесписьменным языкам с ярко выраженным диалектным членением (таким, например, как селькупский) или применительно к корпусу древних письменных памятников (например, древненовгородских).

Добавим также, что ряд искажений и аномальных форм, регулярно встречающихся в Корпусе, объясняется тем, что пишущие

используют так называемую «речевую маску» как прием языковой игры [Земская 1973; Гловинская 1996; Санников 1999], например:

[Дама в фиолетовом]. *И старушка Изергиль с ними?*

[Дама в синем]. *А як же ж! Глянь, кто это там на кухне посудку намывает?*
*[Марина Палей. Long distance, или славянский акцент]*

Существуют конвенционализованные речевые маски, правила употребления которых, безусловно, следует включать в полное описание современного русского языка. Наиболее распространенные среди них — маска «рязанского мужика», которую можно опознать по словам типа *чаво*, *таперича*, а также восточнославянская (*усё*, *як, повбивав бы*), кавказская (*дэвушка*, *канэшна дарагой*, *пачиму*), эстонская и др. Если добавить к этому унаследованные современным русским языком «наслоения» из церковнославянизмов и других архаических оборотов (*Возвращается ветер на круги своя*; *три дни*), а также из диалектной речи (*семь суток не спамши*), то окажется, что современный русский язык не имеет четких границ — ни в пространстве (поскольку отражает диалектные и иноязычные вкрапления), ни во времени (поскольку отражает церковнославянизмы и «застывшие» старые формы); нет четкой грани между письменной и устной речью (в той степени, насколько особенности устной речи фиксируется в письменной).

Таким образом, русская морфология с точки зрения корпуса — более «либеральная» и более широкая морфология, чем та, которая представлена в нормативных грамматиках. В теоретической лингвистике на подобные явления обращают внимание сравнительно редко (исключением являются исследования по русской разговорной речи, начатые еще в 1960-е гг. по инициативе М. В. Панова и Е. А. Земской, ср. [Земская 1973] и др., а также, например, недавние исследования [Гловинская 1996, 2001], содержащие очень показательный в этом отношении материал). Аналогичные проблемы возникают, естественно, даже в большем количестве и при разметке корпусов устной речи. Более широкий учет подобных особенностей является самой актуальной ближайшей задачей развития системы морфологической разметки и расширения грамматического словаря.

**Литература**

Буслаев Ф. И. Историческая грамматика русского языка. М., 1959.

Вайс 1993 — Вайс Д. Двойные глаголы в современном русском языке // Категория сказуемого в славянских языках: модальность и актуализация (Акты международной конференции). München: Sagner, 1993. С. 67-97.

Виноградов В. В. Русский язык. 2-е изд. М., 1972.

Герд, Захаров 2004 — Герд А. С., Захаров В. П. Нерешенные вопросы национального корпуса русского языка // Международная конференция «Корпусная лингвистика — 2004». Тезисы докладов. СПб.: СПбГУ, 2004. С. 28-29.

Гловинская М. Я. Активные процессы в грамматике (на материале инноваций и массовых языковых ошибок) // Русский язык конца XX столетия (1985-1995). М., 1996, с. 237-305.

Гловинская М. Я. Активные процессы в грамматике (на материале инноваций и массовых языковых ошибок) // Земская Е. А. (ред.) Язык русского зарубежья: общие процессы и речевые портреты. М. — Вена, 2001, с. 341-492.

Еськова Н. А. Сведения о грамматических формах // [Борунова С. Н., Воронцова В. Л., Еськова Н. А.] Орфоэпический словарь русского языка: произношение, ударение, грамматические нормы. М., 1983.

Зализняк 1967 — Зализняк А. А. Русское именное словоизменение. М.: Наука, 1967.

Зализняк 1977/2003 — Зализняк А. А. Грамматический словарь русского языка. 1-е изд. М.: Русский язык, 1977 / 4-е изд. М.: Русские словари, 2003.

Земская Е. А. (ред.) Русская разговорная речь. М.: Наука, 1973.

Копотев 2004 — Копотев М. «Несмотря на» «потому что», или Многокомпонентные единицы в аннотированном корпусе русских текстов // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2004. М., 2004.

Копотев М., Мустайоки А. Принципы создания Хельсинкского аннотированного корпуса русских текстов (ХАНКО) в сети Интернет // Научно-техническая информация. Сер. 2: Информационные системы и процессы. № 6: Корпусная лингвистика в России, 2003. С. 33-37.

Кузнецов 2002 — Современный толковый словарь русского языка. СПб.: Норинт, 2002.

Мельчук 1995 — Мельчук И. А. Русский язык в модели «Смысл ⇔ Текст». М.: Языки русской культуры, 1995.

Ожегов, Шведова, 1999 — Ожегов С. И., Шведова Н. Ю. Толковый словарь русского языка. 4-е изд. М.: Азбуковник, 1999.

Плунгян В. А., Сичинава Д. В. Морфологическая информация в Национальном корпусе русского языка // II Международный конгресс исследователей русского языка «Русский язык: исторические судьбы и

современность». Москва, 18-21 марта 2004 г. Труды и материалы. М.: МГУ, 2004.

Грамматика 1980 — Русская грамматика. М.: Наука, 1980.

Санников В. З. Русский язык в зеркале языковой игры. М., 1999.

Чельцова Л. К. Форма множественного числа как объект лексикографии. Дисс... канд. филол. наук. М., 1976.

Шмелева Е. Я., Шмелев А. Д. «Неисконная русская речь» в восприятии русских // Логический анализ языка. Образ человека в культуре и языке. М., 1999.

Corbett G. G. The head of Russian numeral expressions // Greville G. Corbett, Norman M. Fraser and Scott McGlashan (eds.), Heads in Grammatical Theory. Cambridge: Cambridge University Press, 1993. P. 11-35.

Miller 1970 — Miller J. Stative verbs in Russian // Foundations of language, v. 6.4. 1970.