

Д. В. Сичинава

ОБРАБОТКА ТЕКСТОВ С ГРАММАТИЧЕСКОЙ РАЗМЕТКОЙ: ИНСТРУКЦИЯ РАЗМЕТЧИКА

0. ВВОДНЫЕ ЗАМЕЧАНИЯ

Нижеследующий текст, начиная с п. 1, представляет собой несколько сокращённую публикацию реально используемой инструкции по ручному снятию грамматической омонимии (и иным видам редактирования результатов автоматической морфологической разметки) в текстах Национального корпуса русского языка. Инструкцией этой руководствуется «первичный» редактор разметки (в дальнейшем — *разметчик*), результат работы которого затем проверяет и корректирует более опытный эксперт (в дальнейшем — *супервизор*; термин, может быть, не самый удачный, но «исторически сложившийся»).

Первая версия этой инструкции составлена мной в 2002 году на основе аналогичной инструкции, написанной мной совместно с А. Е. Поляковым и использованной в «пилотном» корпусе на основе программы *Mystem* в 2001—2002 гг.¹ В дальнейшем (2002—2005) она неоднократно дорабатывалась на основании решений, принимавшихся на семинарах в ВИНТИ и Институте русского языка им. В. В. Виноградова, так что все участники данных семинаров могут считаться в той или иной мере её соавторами. **Сокращению** подверглись длинные списки конкретных словоформ и словосочетаний с указаниями по их разбору и примерами (полностью даётся только раздел 5.7. — «Количественные наречия и числительные» — составленный Г. И. Кустовой), а также детали, связанные с элементами автоматического снятия омонимии при помощи синтаксического модуля программы *Dialing* и коррекцией соответствующих ошибок. Но, кроме того, **добавлены** некоторые пояснения к тексту в виде подстрочных примечаний: такой жанр дискурса, как инструкция, не предполагает разъяснения того, «зачем» делается тот или иной шаг или «почему», наоборот, не осуществляется некоторая другая операция (между тем как у пользовате-

¹ О том, что это за корпус и по каким принципам он размечался, см. мою статью «Национальный корпус русского языка: очерк предыстории» в настоящем сборнике.

лей Корпуса и читателей его описания такие вопросы закономерно могут и даже должны возникнуть).

1. Постановка задачи

Грамматический анализатор (Dialing_Processor.exe + dial1.bat) порождает для каждой словоформы некоторое **подмножество** всех возможных вариантов разбора. Остальные варианты разбора, исходя из синтаксического контекста, уничтожаются на стадии разбора. Редактору необходимо в каждом случае остающейся неоднозначности выбрать **один** вариант разбора, основываясь на синтаксической структуре и семантике фразы (в некоторых сложных случаях нужно оставлять **несколько** вариантов — п. 4.1). Разбор основан на грамматическом стандарте, который зафиксирован в «Грамматическом словаре русского языка» А. А. Зализняка. К этому словарю и следует обращаться в спорных случаях, не предусмотренных настоящей инструкцией.

Единственный неправильный разбор может возникнуть из-за отсутствия слова в словаре анализатора (см. п. 4.7). Стопроцентно надёжных способов редактирования таких мест не существует.

2. ОБЩИЕ СВЕДЕНИЯ О РАБОТЕ В РЕДАКТОРЕ

2.1. Формат исходного файла

При разборе текст разбивается на квази-слова (tokens), которые бывают двух типов:

- 1) русские словоформы, проходящие грамматический анализ;
- 2) неанализируемые цепочки символов — цифры², некириллические буквы, знаки препинания, а также команды разметки в угловых скобках < >.

За каждой русской словоформой следуют варианты грамматического разбора. Варианты заключаются в фигурные скобки { }, а внутри скобок отдельные варианты разделяются знаком |. Каждый вариант состоит из следующих полей, разделенных знаком =:

- 1) **лексема** (словарная форма);
- 2) **грамматические признаки лексемы** (часть речи и т. д.);
- 3) **грамматические признаки словоформы** (время, число, падеж и т. д.).

² В окончательной версии корпуса цифры получают автоматическую разметку.

Вариант может иметь особенность, помеченную знаком «?» после лексемы. Это значит, что вариант отсутствует в словаре, а сгенерирован программой по аналогии (гипотеза) и может быть неправильным; его можно оставить как есть, исправить или удалить.

В процессе редактирования варианты могут получить дополнительные признаки, которые обозначают другие особенности: грамматические, орфографические и т. д. (см. п. 4).

2.2. Правила работы в редакторе

Программа GRAMEDIT.DOT представляет собой простой редактор для грамматически размеченных текстов, который обеспечивает среду для просмотра и выбора вариантов разбора. Редактор реализован как шаблон для Microsoft Word.

Для удобства работы в редакторе разные элементы текста выделяются разными стилями:

1) варианты разбора в { } и команды разметки < > оформляются как скрытый текст и обычно не видны;

2) словоформы, имеющие более 1 варианта разбора, оформляются синим цветом. Так же оформляются и словоформы, имеющие единственный гипотетический разбор.

Для раскраски текста выберите из меню «Gamedit» пункт «Раскрасить». Для переключения показа скрытого текста выберите из меню пункт «Показать скрытый».

Для редактирования текста выберите из меню «Gamedit» пункт «Редактировать». Вы увидите диалоговое окно, в котором есть список вариантов и набор кнопок. Выберите нужный вариант и нажмите кнопку «Заменить» или клавишу Enter. Если Вы не уверены в своем выборе, оставьте все как есть и нажмите кнопку «Найти >>» или «Найти <<». Если Вы сделали ошибку, нажмите кнопку «Отменить» и затем «Найти <<» для возврата к предыдущему месту. Для выхода из диалога нажмите кнопку «Cancel» или клавишу Escape. Вы можете перемещаться по тексту клавишами курсора, **но для редактирования всегда вызывайте пункт «Редактировать» из меню; редактировать текст напрямую разрешается только в особых случаях** (см. пп. 4.4-4.7).

Если Вы уверены, что данная словоформа (цепочка букв) **всегда** имеет **единственный** вариант разбора, можно нажать кнопку «Заменить все», однако это небезопасно. Для отмены нажмите клавишу «Отменить» два раза.

Грамматическая разметка в Корпусе: Инструкция

Варианты, сгенерированные программой (со знаком ? после лексем), могут быть **все неправильными**. Выберите наиболее подходящий вариант и отредактируйте его при помощи кнопки «Исправить», а затем нажмите кнопку «Заменить». Если ни один из вариантов не подходит, можно нажать кнопку «Удалить», но тогда правильный вариант придется ввести вручную.

3. УСЛОВНЫЕ ОБОЗНАЧЕНИЯ ГРАММАТИЧЕСКИХ КАТЕГОРИЙ

Грамматические категории кодируются следующим образом¹:

1) часть речи:

S	Существительное
A	Прилагательное
NUM	Числительное
V	Глагол
ADV	Наречие
PRAEDIC	Предикатив
S-PRO	Местоимение-существительное (я)
A-PRO	Местоимение-прилагательное (<i>который</i>)
A-NUM	Числительное-прилагательное (<i>шестой</i>)
PRAEDIC-PRO	Местоимение-предикатив (<i>некого, нечего</i>)
PARENTH	Вводное слово
PR	Предлог
CONJ	Союз
PART	Частица
INTJ	Междометие ²

2) именные категории:

род:	m, f, n
одушевленность:	anim, inan
число:	sg, pl
падеж:	nom, gen, dat, acc, ins, loc, loc2 (второй предложный: <i>в лесу</i>), gen2 (второй родительный: <i>чашка чаю</i>), adnum (счётная форма: <i>два часа</i>); вручную вводится также падежная форма acc2 (второй винительный: <i>метит в генералы</i> ; см. об этой форме ниже подробнее, п. 4.10), voc (звательный падеж — см. там же, п. 4.11).

¹ Во «внутреннем формате» редактируемой разметки вместо помет на латинице до недавнего времени использовались пометы на кириллице (порождаемые программой Dialing); они однозначно соответствуют принятым в корпусе латинским пометам, и это не создаёт сложностей.

² В этом списке отсутствует местоименное наречие (ADV-PRO), которое вносится в корпус автоматически по словарю (соответствующие словоформы разбираются во внутреннем формате как ADV).

Д. В. Сичинава

краткость: **brev**³
степень сравнения: **comp, comp2** (форма с приставкой *по-*)

3) глагольные категории:

вид: **pf, ipf**
переходность: **tran, intr**
залог: **act, pass**⁴
репрезентация: **inf, partcp, ger**
наклонение: **imper, imper2** (форма на *-мте: пойдёмте*)⁵
время: **praes, praet, fut**
лицо: **1p, 2p, 3p**

4) особые пометы

нестандартная форма: ***6**
искаженная форма: **!7**
запись цифрами: **ci^{ph}**
сокращение: **abbr**
имя: **persn**
фамилия: **famn**

Примечания:

а) у сравнительных форм (компаративов) частеречный признак указывается дизъюнктивно (A/ADV), лексема при этом указывается адъективная, например, *выше*{высокий=A/ADV= comp}

б) неизменяемые местоимения-прилагательные *его, её, их* не имеют словоизменительных характеристик;

в) несклоняемые существительные и прилагательные имеют особую помету 0 (ноль): *шоссе*{шоссе=S,n,inan,0=sg,nom|...}

4. ОСНОВНЫЕ ПРОБЛЕМЫ ПРИ ВЫБОРЕ ВАРИАНТОВ

4.1. Сгенерированные варианты

Уберите все заведомо бессмысленные варианты (глагол вместо существительного и т. п.) и отредактируйте наиболее правдоподобный вариант.

³ Помета **plen** (полная форма) в дальнейшем вносится автоматически.

⁴ Страдательный залог усматривается только у причастий. Помета **med** (средний залог) для форм на *-ся* вносится автоматически.

⁵ Помета **indic** (изъявительное наклонение) в дальнейшем вносится автоматически.

⁶ В дальнейшем автоматически меняется на **apom**; обозначение одним символом выбрано для удобства разметчика.

⁷ То же; в дальнейшем автоматически меняется на **distort**.

Грамматическая разметка в Корпусе: Инструкция

Обращайте внимание на род и одушевленность существительных! Учтите, что

имена людей должны иметь признак

S,persn,m,anim или S,persn,f,anim,

фамилии —

S,famn,m,anim или S,famn,f,anim.

Мужская и женская фамилии считаются отдельными лексемами.

Множественное число от фамилий (*Домбровские, Петровы*) разбирается как множественное число от «мужского» варианта.

Отчества не имеют особой пометы и записываются просто как S,m,anim или S,f,anim⁸.

Географические названия разбираются как *singularia tantum*, с пометой S,m,inan,sg, S,f,inan,sg или S,n,inan,sg (если речь не идёт о топонимах *pluralia tantum* вроде *Люберцы* — они получают разбор S,pl — или о маркированных употреблениях вроде *две Германии* или *галопом по Европам*).

У аббревиатур число указывается в признаках словоформы или лексемы в зависимости от сингулярности или множественности означаемого ею объекта:

ООН{ООН=S,f,inan,sg,0=nom}; ЖЭК{ЖЭК=S,f,inan=sg,nom};

род соответствует роду их главного слова (если это не противоречит контексту):

вручают премию в ООН{ООН=S,f,inan,sg=loc} (*организация*); НО

наш ЖЭК{ЖЭК=S,m,inan=sg,nom} (*хотя контора*).

Сокращения следует раскрывать, сопровождая разбор пометой «=abbr»:

т{товарищ=S,m,anim=m,sg=abbr},,

г{господин=S,m,anim=m,sg=abbr},,

см{смотреть=V,ipf,tran=act,imper,2p,sg=abbr}.

Инициалы (*Л. Н. Толстой*) автоматически получают разбор вида {Л=INIT=abbr}; если разбор ошибочен (например, когда инициал совпадает с однобуквенным словом или сокращением), его следует поправить.

⁸ В дальнейшем помета *patr*n (отчество) вносится в корпус по словарю отчеств автоматически.

Д. В. Сичинава

Словоформы, которые частично записаны цифрами, получают разбор, лексема в котором записана так же, как и в словоформе, а в конце разбора стоит помета ciph. Например:

17-м{17-й=A-NUM=n,sg,ins=ciph},
225-летие{225-летие=S,n,anim=sg,nom=ciph}

Если Вы откорректировали вариант и/или удалили вопросительный знак после лексем, поставьте прямо после лексем знак * (звездочка) — это означает, что такая лексема отсутствует в словаре морфологического анализатора⁹, например:

{Грибоедов*=S,фам,m,anim=sg,nom}
{Тифлис*=S,m,inan,sg=nom}

Если слово, отсутствующее в словаре, несклоняемое, после признаков лексем ставится знак 0 (ноль):

{Брижит*=S,имя,f,anim,0=sg,nom} {Бардо*=S,фам,f,anim,0=sg,nom}.

Слова, отсутствующие в словаре и порожденные в качестве гипотез, как правило, имеют неупорядоченный набор граммем. Старайтесь, редактируя правильный вариант разбора, распределять граммемы на граммемы лексем и словоформы, а внутри этих последних групп придерживаться порядка, принятого в разборах слов, наличествующих в словаре.

Например, программа даёт:

мурчанием{мурчание?=S,n,sg,inan=ins}

Вы пишете:

мурчанием{мурчание*=S,n,inan=sg,ins}

Если Вы не уверены, как записать правильный вариант, лучше оставьте все как есть.

NB Нужно оставлять несколько вариантов:

- а) в случае неразрешимой омонимии родительного и винительного при отрицании (*не знал родного отца*);
- б) в случае контекстно неразрешимой одушевленности/неодушевленности (*манекен, кукла*; но: *я не люблю манекенов* — anim, *я не люблю манекены* — inan)

⁹ Звёздочка после лексем используется для пополнения словаря программы (в него добавляются наиболее частотные в корпусе слова, отсутствующие в словаре). В корпус этот символ не переносится.

в) в случае, если форма лексемы колеблется, а употребленная косвенная форма не позволяет это различить (*спазмами* — от *спазм* или *спазма*?).

4.2. Прописная и строчная буквы в лексеме

Большая (прописная, заглавная) буква в лексеме имеется только: **а)** у всех имен собственных, означающих живые существа: *Максим*{Максим=S,persn,m,anim=sg,nom}, *Васильев, Джон, Шарик*, **б)** у топонимов: *Ленинград*{Ленинград=S,m,inan,sg=nom}, *Великие Луки*; **в)** у имен классов **а)** и **б)**, употребляемых как имена собственные различных других объектов: «*Прага*», «*Максим Горький*». Если словарь даёт такую лексему со строчной буквы, следует править строчную букву на заглавную, звёздочку при этом не ставить.

Притяжательные прилагательные от имён собственных пишутся в поле лексемы со строчной буквы, даже если в тексте они с прописной: *Машина*{машин=A=f,sg,nom} *книга*; *Петров*{петров=A=m,sg,nom} *день*. Это же относится и к словам *Бог, Господь, Богородица, Государь, Цесаревич* и тому под.; соответствующие лексемы пишутся со строчной буквы (*бог, господь...*) вне зависимости от орфографии автора¹⁰. Строчная буква сохраняется в лексеме и у нарицательных существительных, выступающих в функции собственного имени неодушевленных предметов («*Правда*{правда=S,f,inan=sg,nom}», «*Ударник*»); названия марок вроде *волга, шкода, форд*, имеющиеся в словаре Зализняка, также даются со строчной буквы.

4.3. Нестандартные формы

Грамматическая форма может быть распознана неверно из-за особенностей авторского языка. Если эта форма не соответствует правилам современного русского языка (согласно словарю Зализняка), укажите правильный набор грамматических признаков, но поставьте после него знак * (звездочка), что означает нестандартную форму, например:

¹⁰ Это делается ради единого учёта соответствующих лексем в (будущем) слово-указателе к корпусу. В текстах представлены значительные колебания, в случае с сакральной лексикой особенно заметные, разумеется, при сравнении официально опубликованных советских текстов, с одной стороны, и несоветских и постсоветских — с другой (вторая группа текстов, впрочем, тоже обнаруживает явную пестроту).

Д. В. Сичинава

<i>три дни</i>	{день=S,m,inan=sg,gen*}
<i>для доклада</i>	{доклад=S,m,inan=sg,gen*}
<i>кудахтая</i>	{кудахтать=V,ipf,intr,act=ger,praes*}

4.4. Авторское написание

Авторское написание используется для передачи иностранного акцента (*дэвушка*) или других особенностей произношения (*де-е-е-вуш-ка*). Мы сохраняем авторское написание, но даем к нему **правильный** вариант разбора, который можно взять из другого места в тексте или исправить имеющийся. Орфографическая особенность помечается знаком «=!» в конце, например:

<i>дэвушка</i>	{девушка=S,f,anim=sg,nom=!}
<i>де-е-е-вуш-ка</i>	{девушка=S,f,anim=sg,nom=!}

Особую проблему представляют междометия, например *а-а-а*, *э-э*, *м-м-ммм*, *фу-у*. Если у междометия есть основная, фиксированная в словаре Зализняка форма (напр. *фу*, *батюшки*, *мм*), то такая запись разбирается как произносительный вариант стандартной формы {фу=INTJ=!}, в противном случае возможна запись *а-а* и *э-э* как отдельных лексем (различное значение: *Э!* — оклик, *э-э* — подбор слова).

4.5. Иноязычные вставки

Разбор иноязычных вставок на латинице нужно обязательно удалить. Иногда попадают иноязычные вставки, записанные кириллицей, которые получают какой-то разбор. Если это явно иноязычный текст, даже на близко родственном языке (белорусском, украинском), то разбор нужно удалить целиком, включая фигурные скобки и синтаксические пометы. Если это все-таки русский текст с диалектными вставками, то разбор можно сохранить, пометив лексему или грамматическую форму знаком * (элемент авторского языка). Так же следует поступать и с церковнославянскими лексемами и грамматическими формами вроде

<i>Возвращается ветер на кругу своя</i>	{свой=A-PRO=pl,acc*};
<i>Блажен муж, иже муж не иде</i>	{идти=V,ipf,intr,act=sg,praet*}
<i>на совет нечестивых</i>	

Сложные случаи решает супервизор.

4.6. Слова с дефисом

Слова с дефисом почти всегда анализируются программой как единые лексемы типа *баба-яга*, *генерал-полковник*, с одним разбо-

Грамматическая разметка в Корпусе: Инструкция

ром после второй половины, хотя на самом деле это в большинстве случаев свободные сочетания двух лексем. Сочетания бывают следующих типов:

- а) приложения (зять-машинист, Вовка-критик);
- б) аппроксимативные конструкции вида числительное-числительное (**две-три** книги);
- в) удвоенные наречия (долго-долго)
- г) различной структуры и семантики глагольные конструкции (постояй-постояй; **молчит-молчит**, да вдруг как скажет)

Впрочем, в большинстве случаев конструкции этого последнего типа разбираются вполне верно.

Например, разборы:

зять-машинист{зять-машинист?=S,m,anim=sg,nom}
Вовка-критик{Вовка-критик?=S,m,anim=sg,nom}
две-три{два-три?=NUM=nom|два-три?=NUM=f,nom|два-три?=NUM=acc|два-три?=NUM=f,acc} книги
долго-долго{долго-долго?=ADV}
постояй-постояй{постояй-постояй?=S=m,sg,nom,inan|постояй-постояй?=S=m,sg,acc,inan}

должны превратиться в:

зять{зять=S,m,anim=sg,nom}-*машинист*{машинист=S,m,anim=sg,nom}
Вовка{Вовка*=S, имя,m,anim=sg,nom}-*критик*{критик=S,m,anim=sg,nom}.
две{два=NUM=f,nom}-*три*{три=NUM=nom} книги
долго{долго=ADV}-*долго*{долго=ADV}
постояй{постоять=V,pf,intr,act=sg,2p,imper}-
постояй{постоять=V,pf,intr,act=sg,2p,imper}

Сложные имена собственные, прилагательные, а также лексикализованные сочетания сохраняют дефис в лексеме:

Экс-ан-Прованс{Экс-ан-Прованс*=S,m,inan,sg=nom}
Мария-Терезия{Мария-Терезия*=S,persn,f,anim=sg,nom}
серо-буро-малиновый{серо-буро-малиновый*=A=m,sg,nom}
непонятно-бледный{непонятно-бледный*=A=m,sg,nom}
плащ-палатка{плащ-палатка=S,m,anim=sg,nom}

О случаях, когда дефис передаёт особенность произношения (*де-е-е-вушка*, *вы-хо-ди*), см. выше, п. 4.4.

4.7. Единственный вариант

Иногда из-за той или иной лакуны в словаре программа порождает единственный вариант разбора, который не годится, например:

Д. В. Сичинава

Грибов{гриб=S,m,inan=pl,gen}

— на самом деле это фамилия *Грибов*.

Груша{груша=S,f,inan=sg,nom}

— на самом деле это имя *Груша*.

покрупней{покрупнеть=V,pf,intr,act=sg,2p,imper}

— на самом деле это особая сравнительная степень от *крупный*.

Поскольку редактор пропускает однозначные разборы, то обнаружить такие случаи можно только путем визуального просмотра. В таком случае нужно показать скрытый текст и исправить разбор вручную, например, в нашем случае, на

{Грибов*=S,famn,m,anim=sg,nom},

{Груша*=S,persn,f,anim=sg,nom},

{крупный=A/ADV=comp2}.

Все такие случаи по мере обнаружения надлежит выписывать в особый файл и предоставлять его супервизору вместе с размеченным текстом. Список будет использован при усовершенствовании словаря.

4.8. Слова на пол-

Отсутствующие в словаре программы существительные вида «*пол*+счетная форма» (но, разумеется, не «*пол*+Им.» — «полдень», «полночь» и тому под.), независимо от орфографии, надо разбирать так¹¹:

пол-арбуза{пол=NUM+арбуз=S,m,inan=sg,gen}

полгаллона{пол=NUM+галлон=S,m,inan=sg,gen}

пол-Европы{пол=NUM+Европа=S,f,inan,sg=gen}

К формам косвенных падежей (*получаса*, *полумира*, *полугаллона*) это не относится: их следует, согласно словарю Зализняка, разбирать как обычные формы от лексем *получас*, *полумир*, *полугаллон*.

4.9. «Не», «полу» и тому подобное

Слова, образованные соединением приставки или наречного компонента *P* с причастием от глагола *V*, если слитно пишущийся глагол *P+V* в русском языке заведомо не существует, записываются так:

неопохмелившийся{не=PART+опохмелиться=V,pf,intr,act=partcp,praet,m,sg,nom}

свежевылитый{свеже+вылить=V,pf,tran=partcp,praet,m,sg,nom,pass}

¹¹ Тем самым усматривается особая лексема: числительное *пол*.

Но:

недокрашенный{недокрасить*=V,pf,tran=partcp,praet,m,sg,nom,pass}

Прочие слова с «не», «полу» и тому под. надо разбирать как новые лексемы:

нечеченец{нечеченец*=S,m,anim=sg,nom}

некрутой{некрутой*=A=m,sg,nom}

полусолдат{полусолдат*=S,m,anim=sg,nom}

полуобразованный{полуобразованный*=A=m,sg,nom}

4.10. Второй винительный падеж

После предлога **в** именительный падеж одушевленных имен помечается вопросительным знаком:

годится в отцы{??|отец=S,m,anim=pl,nom}

Исправьте, если нужно, именительный падеж на второй винительный:

{отец=S,m,anim=pl,acc2}

Второй винительный падеж усматривается¹² также у синтаксически управляемых форм числительных, совпадающих с формой именительного падежа и указывающих на количество объектов, означаемых одушевленным существительным:

больше на три{три=NUM=acc2} человека

вызывают по три{три=NUM=acc2} человека

силой в три{три=NUM=acc2} медведя

обслужить двадцать три{три=NUM=acc2} клиента.

При неодушевленных объектах даётся простой винительный падеж:

больше на три{три=NUM=acc} метра

наносят по три{три=NUM=acc} удара и т. д.

4.11. Некоторые случаи второго предложного падежа

Обратите внимание на то, что второй предложный падеж в ряде случаев отличается от предложного только ударением: *невиновен в крóви* — *храм на кровí*, *говорит о пéчи* — *лежать на пещí*; программа в таких случаях предлагает вариант «loc2».

¹² Ср. у И. А. Мельчука анализ случаев вроде *силой в три медведя* как форм винительного падежа (обычного, не «второго» или «включительного») с утратой одушевлённости: *Русский язык в модели Смысл—Текст*. М.—Вена, ЯРК, 1995: 515—536; аналогичный анализ конструкции *пойти в солдаты* там же, 537—563.

4.12. Счетная форма

Для четырёх русских существительных — *час*, *шаг*, *шар* и *ряд* — существует так называемая счётная форма (**adnum**), употребляемая при сочетании с числительными *полтора*, *два*, *три* и *четыре*, а также с существительным *четверть*. Она отличается от родительного падежа единственного числа местом ударения: *часá*, *шагá*, *шарá*, *рядá*.

При сочетании лексемы *час* с числительными *полтора*, *два*, *три* и *четыре* в качестве единственного варианта даётся **adnum** (т. е. *часá*). Словоформа *полчасá* разбирается как

{пол=NUM+час=S,m,anim=sg,adnum}

При сочетании лексемы *час* с существительным *четверть*, а также лексем *шаг*, *шар* и *ряд* со всеми названными словами следует оставлять **оба** варианта: **gen** и **adnum**.

4.13. Звательная форма

Звательная форма, как старая, так и новая, помечается «voc»:

отче{отец=S,m,anim=sg,voc}

Петь{Петя=S,persn,m,anim=sg,voc}

Слова *боже* и *господи* в словаре А. А. Зализняка значатся как междометия. Рекомендуется исправлять разборы на

{бог=S,m,anim=sg,voc}

и

{господь=S,m,anim=sg,voc}

в контекстах, когда мы имеем дело с явными обращениями к Богу: *... но Господи, Боже мой, ты же знаешь, я никогда не любил ее, такую безобразную, грубую, плечистую — ни дать ни взять, переодетый мельник* (Домбровский), а также в случаях, когда с ними согласованы определения: *Боже мой, милостивый Господи*.

На всякий случай специально отметим, что если именительный падеж употребляется в звательной функции, это никак не отмечается: *Богородица царевна*{царевна=S,f,anim=sg,nom} *Киргиз-Кайсацкия орды*, <...> *подай найти её совет*.

4.14. Родительный при отрицании

Иногда при отрицании формальные признаки не позволяют различить родительный и винительный падежи: у несклоняемых слов (*не слушал радио*; ср. *не смотрел телевидение/телевидения*), у

местоимений (*он читал книгу, а я её не читал; ср. не читал книгу/книги*) и прежде всего у одушевленных существительных и определяющих их прилагательных и причастий (*он не знал родного отца; ср. не знал родную мать/родной матери*). Во всех таких случаях должна оставаться неснятая омонимия.

4.15. Формы с -(е/а)йш- и наи-

Форма с суффиксом *-(е/а)йш-* (*умнейший, высочайший*) считается не превосходной формой прилагательного, а самостоятельной лексемой. Также самостоятельными лексемами считаются все прилагательные с приставкой *наи-* (*наикратчайший, наиредчайший*).

5. НАИБОЛЕЕ ВАЖНЫЕ СЛУЧАИ ОМОНИМИИ

5.1. Наречия, краткие прилагательные, предикативы

В русском языке существует масштабная регулярная омонимия (синтаксическая полифункциональность) форм на **-о/-е**, а именно, они могут быть: а) наречиями б) краткими формами прилагательного в) предикативами. В разметке все эти три функции принято различать следующим образом:

Наречие определяет глагол, прилагательное или иное наречие: он *неприлично* ругается, он *неприлично* откровенен, он приходит *неприлично* редко.

Краткое прилагательное является частью составного именного сказуемого при подлежащем-существительном (местоимении) среднего рода или инфинитиве: это было *хорошо*, уходить не прощавшись — *неприлично*, догадаться было *легко*.

Предикатив также является сказуемым в предложении и тоже может присоединять слова *было, будет*, но он, в отличие от краткого прилагательного и тем более от наречия, может управлять:

а) субъектом в дательном падеже: мне было *хорошо*;

б) синтаксическим дополнением, финитным или нефинитным: ей было *жутко* ломать крышу, мне было *больно*, что начнут отрывать доски; *хорошо*, что она так думала.

В сложных случаях оставляйте неснятую омонимию.

Замечание. В словаре Зализняка отсутствует огромное количество отадективных наречий (не говоря уже об омонимичных предикативах), например: *забавно, нечаянно, прелестно, проворно, уныло, старательно, торопливо* и т. д. Все такие формы получают един-

Д. В. Сичинава

ственный вариант разбора «A=brév,n». Фильтр помечает такие случаи знаком вопроса; если Вы уверены, что перед Вами наречие или предикатив, нужно вводить вручную разбор вроде забавно*=ADV, забавно*=PRAEDIC.

5.2. Сравнительная степень

Как уже сказано выше (п. 2, список грамем), не различаются формы сравнительной степени от прилагательных и отадъективных наречий. Все такие формы получают комбинированный признак «прилагательное/наречие»¹³, например:

лучше {хороший=A/ADV=comp}.

В словаре Зализняка отсутствуют формы сравнительной степени с приставкой *по-*: *побольше, повыше, поскорее* и т. д. В таких случаях замените сгенерированный вариант на исходное прилагательное, а в грамматических характеристиках укажите comp2, например:

повыше {высокий=A/ADV=comp2}.

5.3. Творительный падеж и наречие

В нашем разборе исключены наречия, омонимичные с формой творительного падежа существительного, например, *порой, месяцами, ночью, гужом* — прежде всего потому, что они сочетаются с прилагательными в творительном падеже без изменения семантики: *весенней порой, поздней ночью*. Тем не менее оставлены некоторые наречия, расходящиеся с именем акцентно и/или семантически: *бегóm, верхóm, даром, кругóm, навалом, особняком, порядком, рядком, рядом, следом, толком, часом*.

5.4. Причастия и прилагательные

Прилагательные, производные от причастий: *открытый, согнутый, вооруженный, дисциплинированный* и т. д. — омонимичны исходным причастиям, отделить одни от других, семантически или синтаксически, сложно. В таких случаях почти всегда следует оставлять неснятую омонимию (удаляя вручную только неверные падежные варианты). Исключения бывают в следующих случаях:

1) прилагательное образовано от совсем другого слова и/или исходное причастие мало употребительно:

сметанный (от сметана) — *смётанный*

¹³ Это соответствует точке зрения, согласно которой «компаратив» представляет собой особенную часть речи.

Грамматическая разметка в Корпусе: Инструкция

тяжеленный (очень тяжелый) — причастие от *тяжелить* (?)

численный (от число) — причастие от *числить* (?)

мысленный (от мысль) — причастие от *мыслить* (?)

2) прилагательное и причастие различаются по форме (ударению, ё), например:

совершенный — *совершённый*

приближённый — *приближенный*.

5.5. Счетная форма существительного.

Числительные 2-4 в именительном падеже управляют формой «sg,gen»¹⁴, а не «pl,nom», например: *две собаки, три тетради, четыре правила*.

5.6. Частицы и союзы

Некоторые слова могут быть и частицами, и союзами. Союз вводит целое предложение и стоит, как правило, в его начале. Союз несёт дополнительное значение (изъяснительное, противительное...), которое частице не свойственно. Сфера действия частицы — лишь часть предложения.

Вот список некоторых таких слов: *будто, ведь, даже, же, ли, лишь, пусть, ровно, словно, точно, хоть, якобы*¹⁵.

5.7. Количественные наречия и числительные: мало, много, меньше, больше...

МНОГО = ЧИСЛИТЕЛЬНОЕ:

подлежащее: *много лет прошло* — {много=NUM=nom}

дополнение: *много хочешь; много знает* — {много=NUM=acc}

обстоятельство меры: *видел много раз, прошел много миль, много лет прожил в Магадане* — {много=NUM=acc}

МНОГО = НАРЕЧИЕ:

Много {много=ADV} разговариваешь!

Шумишь много {много=ADV}, а толку чуть!

МАЛО = ЧИСЛИТЕЛЬНОЕ

Мне мало известно (=немногое, почти ничего)

Как мало он помнит {мало=NUM=acc}

¹⁴ Таким образом, счётная форма обособляется в качестве отдельной граммы только в случае акцентного различия с формой родительного падежа (см. п. 4.12).

¹⁵ Подробный разбор неоднозначных слов с примерами опускается.

Д. В. Сичинава

Я прошу/хочу так мало {мало=NUM=acc} = малого, немногого
Мало кто пришел {мало=NUM=nom} (=мало тех, кто); так же:
мало на что влияет, мало {мало=NUM=acc} *что сделал* и т. п.
Мало ли сейчас шляется людей с оружием {мало=NUM=nom}

МНОГО = ПРЕДИКАТИВ

Не наливай, ему много будет {много=PRAEDIC}

В отличие от предикатива МАЛО, предикатив МНОГО почти не встречается.

МАЛО = НАРЕЧИЕ

мало известен (=не очень; наречие);

Люди мало верят в выборы

Из отношения мало зависят от городских властей

МАЛО = ПРЕДИКАТИВ:

Гуманизма в нем мало {мало=PRAEDIC}, ср. *Гуманизма в нем нет совсем*

Хорошего в этом мало

БОЛЬШЕ = СРАВНИТ. СТЕПЕНЬ от ЧИСЛИТЕЛЬНОГО МНОГО

Больше {много=NUM=compr} *всего денег ушло на продукты*

Потратил больше сил, чем вчера — {много=NUM=compr}

Я забыла больше, чем помню — {много=NUM=compr}

Пять рублей больше двух рублей — {много=NUM=compr}

Ей лет двадцать пять, не больше — {много=NUM=compr}

Ванной не пользовались больше года (=много времени) — {много=NUM=compr}

БОЛЬШЕ = СРАВНИТ. СТЕПЕНЬ от ПРИЛАГАТЕЛЬНОГО БОЛЬШОЙ

Пенсия стала больше {большой=A/ADV=compr}

Иногда пенсия больше зарплаты {большой=A/ADV=compr}

Братья всегда хватали куски побольше {большой=A/ADV=compr2}

БОЛЬШЕ/БОЛЕЕ = СРАВНИТ. СТЕПЕНЬ НАРЕЧИЯ МНОГО (заменяется на ОЧЕНЬ, ВЕСЬМА, СИЛЬНО, УЖАСНО)

Испугался козла больше тигра — {много=ADV=compr}

...следует пожалеть больше, чем Иуду — {много=ADV=compr}

Он все больше/более путался в показаниях — {много=ADV=compr}

Более всего на свете прокуратор ненавидел запах розового масла — много=ADV=compr} (= очень)

БОЛЬШЕ = СРАВНИТ. СТЕПЕНЬ от ПРЕДИКАТИВА

Проблем стало больше {много=PRAEDIC=compr}

В январе времени будет побольше {много=PRAEDIC=compr2}

Грамматическая разметка в Корпусе: Инструкция

БОЛЬШЕ/БОЛЕЕ = НАРЕЧИЕ

Я вас больше не задерживаю — {больше=ADV}

Не будем больше загружать телеграф — {больше=ADV}

Говорить более было не о чем — {более=ADV}

МЕНЬШЕ = СРАВНИТ. СТЕПЕНЬ от ЧИСЛИТЕЛЬНОГО МАЛО

Храмы посещает еще меньше народу — {мало=NUM=comp}

Меньше шансов, чем раньше — {мало=NUM=comp}

У нее три пары тапок, не меньше — {мало=NUM=comp}

До выборов остается меньше — {мало=NUM=comp} года

Вложения окупаются меньше чем за год — {мало=NUM=comp} (= за время, которое меньше года)

МЕНЬШЕ = СРАВНИТ. СТЕПЕНЬ от ПРИЛАГАТЕЛЬНОГО МАЛЕНЬКИЙ

Зарплата стала меньше {маленький=A/ADV=comp}

МЕНЬШЕ = СРАВНИТ. СТЕПЕНЬ от НАРЕЧИЯ МАЛО (= почти не)

Должность влияет меньше, чем положение {мало=ADV=comp}

Боятся тигров меньше козлов

Все меньше боялся козлов

Старайся поменьше поднимать тяжести {мало=ADV=comp2}

МЕНЬШЕ/ПОМЕНЬШЕ = СРАВНИТ. СТЕПЕНЬ от ПРЕДИКАТИВА

На кухне бумаг было поменьше {мало=PRAEDIC=comp2}

И чтобы шуму было поменьше! {мало=PRAEDIC=comp2}

НЕМНОГО = ЧИСЛИТЕЛЬНОЕ

Добавь еще немного соли {немного=NUM=acc}

Немного {немного=NUM=nom} денег ушло на елочные игрушки, остальные — на продукты

НЕМНОГО = НАРЕЧИЕ

Немного удивился осведомленности прокуратора {немного=ADV}

Помолчи немного {немного=ADV}

Немного постоял и ушел {немного=ADV}

Просидел немного {немного=ADV} больше {много=ADV=comp}, чем планировал

МНОГИЙ

В абсолютном употреблении и избирательной конструкции — МЕСТОИМЕНИЕ-СУЩЕСТВИТЕЛЬНОЕ:

Многие отказались — {многий=S-PRO=...}

Многие из них отказались — {многий=S-PRO=...}

В функции определения — МЕСТОИМЕНИЕ-ПРИЛАГАТЕЛЬНОЕ:

Многие {многий=A-PRO=...} другие {другой=S-PRO=...} знали об этом

Примечание

В оборотах *тем не менее*, *тем более*, *более или менее*¹⁶, а также в аналитических формах *более высокий*, *менее сильный* БОЛЕЕ и МЕНЕЕ разбираются как НАРЕЧИЯ.

5.8. Отдельные слова¹⁷

NB. Во многих случаях омонимичный союз или частица отличаются от других омонимов устойчивой фразовой безударностью (да—да́, что—что́, это—это́). Кроме того, союз и частица находятся в такой синтаксической позиции, которая исключает глагольное управление и изменение по падежам, и это легко проверить. Попробуйте изменить фразу с сохранением синтаксической структуры, чтобы спорное слово оказалось в другой форме (например, *из-за* другого управляющего слова). Если это невозможно, значит, это не существительное или прилагательное, а союз, частица или другое несклоняемое слово.

В спорных случаях лучше оставьте все как есть или спросите супервизора.

¹⁶ Программа старается автоматически определять сложные лексические единицы (СЛЕ) и приписывать им пословные разборы. Однако иногда синтаксический анализатор приписывает СЛЕ некоторый анализ, который по внутренним правилам анализа имеет приоритет над словарным разбором (например, *может быть*, *то есть*), и нужный вариант приходится выбирать вручную.

¹⁷ Список отдельных неоднозначных слов и правила их разбора опускаются.