А.В. Венцов, Е.В. Грудева, В.Б. Касевич

МОРФОЛОГИЧЕСКАЯ ПРОБЛЕМАТИКА В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЛИТЕРАТУРНОГО ЯЗЫКА

Национальный корпус русского литературного языка представляет собой собрание письменных текстов (художественных, научно-публицистических, публицистических и драматургических), которые отражают период с середины 50-х гг. до нашего времени. Большинство текстов представлено фрагментами.

Важной особенностью корпуса является обязательная акцентуированность всех словоформ текстов, а также систематическое восстановление в правах буквы «ё». Это позволяет избежать ошибок в морфологическом описании словоформ, связанных с омографией; использовать корпус в качестве надёжного обучающего русскому языку средства (в том числе и для иностранцев), обращаться к нему как при решении традиционных лингвистических задач, так и при моделировании процессов порождения и восприятия речи, в автоматическом синтезе речи по тексту. Во всех этих случаях наличие информации об ударении является обязательным.

В отличие от этого, автоматическая (или полуавтоматическая) морфологическая разметка неакцентуированного русского текста затруднена наличием большого числа (более 4 тыс.) омографов; соответственно, не зная места ударения, компьютерная программа часто не может определить, с какой словоформой имеет дело (например, *любИм* или *лЮбим*?), и вынуждена давать все возможные описания для данной словоформы. Наличие проставленного в текстах ударения позволяет снять омонимию, связанную с ударением и неиспользованием буквы «ё», и избежать возможных ошибок в морфологическом описании словоформ.

При таком подходе описание в корпусе омографических форм сразу принимает разный вид для каждого члена пары (тройки) омографов. Например,

во$^{+}$ды{во$^{+}$ды=NNP,0,inan=,pl,nm}
воды$^{+}${вода$^{+}$=NNS,f,inan=,sg,gn}

или

заступи$^{+}$тесь{заступи$^{+}$ться=VV0,prfc,intr=,0,impr,0,pl,2p,0}
засту$^{+}$питесь{заступи$^{+}$ться=VV0,prfc,intr=,0,indc,futr,pl,2p,0}
(Здесь и далее символом «$^{+}$» помечен ударный гласный.)

При выборе общей структуры морфологического описателя за основу была принята структура, создаваемая программой DiaLing, с той лишь разницей, что часть характеристик словоформы из категории грамматических была отнесена к категории «принадлежность к субпарадигме». В итоге для каждой словоформы прежде всего указывается, к какой лексеме она относится; далее следуют морфологические характеристики лексемы (точнее, основного варианта лексемы), затем указывается принадлежность к субпарадигме (например, у глаголов – отнесение к субпарадигме «причастие», «деепричастие», «инфинитив», или «безличное употребление») и – далее – морфологические характеристики данной словоформы. Общее описание имеет следующий вид:

<словоформа>{<лексема>=<частеречный маркер лексемы>,
<грамматич. характеристики лексемы>
=[<принадлежность к субпарадигме>],
<грамматич. характеристики словоформы>}

Для каждого класса описываемых единиц языка количество дескрипторов и порядок их введения при морфологической разметке есть *характеристика постоянная.* Если некоторая характеристика, предусмотренная системой дескрипторов для единиц данного класса, оказывается отсутствующей (невозможной, неприменимой), ее отсутствие фиксируется символьно введением нуля. Так, в сослагательном наклонении или в императиве глаголы не различаются по временам, но эта характеристика считается неотъемлемым признаком глагола, а потому ее отсутствие в описании отмечается нулём.

Например:

ве+рил{ве+рить=VV0,impf,intr=,0,indc,past,sg,0,m}

ве+рил{ве+рить=VV0,impf,intr=,0,sbjn,0,sg,0,m}

иди+{идти+=VV0,impf,intr=,0,impr,0,sg,2p,0}

При выборе системы имен индексов для описания **частеречных классов** и **субпарадигм** мы ориентировались на Национальный Британский корпус (British National Corpus – BNC), в итоге описатели частеречных классов и субпарадигм представляют собой трехсимвольные сочетания в латинице. Для описания **грамматических характеристик** использована в несколько измененном виде система маркеров, принятая в корпусе русского языка, размещенном на сайте Яндекса.

Принятая **система лексико-грамматических классов** по ряду параметров отличается как от традиционных (академических) систем частей речи, так и от систем, используемых в других корпусах русского языка. Так, кроме традиционных частеречных классов (сущ., прилаг., глагол, союзы, предлоги и пр.), введены такие классы, как: вводные слова (типа *коне+чно, быть_мо+жет, по-мо+ему*), аналитические прилагательные (типа *Горбачё+в-* из *Горбачё+в-фо+нд, Интерне+т-* из *Интерне+т-кафе+*), связанные слова (типа *тё+мно-* из *тё+мно-кра+сный, неме+цко-* из *неме+цко-ру+сский*), служебные слова (типа *са+мый* как средство образования аналитического суперлатива, *бо+лее, ме+нее* – как средства образования аналитического компаратива). Последовательно разграничиваются – условно на уровне частеречной характеристики – полнозначные и вспомогательные глаголы; на этом же уровне выделяются безличные глаголы. Тем самым снимается ещё один пласт морфологической омонимии – например, местоимение *са+мый* (*о+н са+мый*) и служебное слово *са+мый* (*са+мый большо+й*); вспомогательный глагол *бы+ть* (*О+н бы+л серьё+зен*) и полнозначный глагол *бы+ть* (*О+н бы+л вчера+ на рабо+те*); личный глагол *рвё+т* и безличный глагол *рвё+т* и т.п.).

В настоящее время в один класс вспомогательных глаголов попадают: глаголы, служащие для образования аналитических форм будущего времени, аналитических форм пассива, связки, а также фазовые глаголы (типа *начина+ть, конча+ть, продолжа+ть*). Стоит особо оговорить необходимость включения в данный класс фазовых глаголов. С семантической точки зрения они – безусловные операторы. У них всегда валентность (единственная) на пропозицию (*начина+ть, конча+ть, продолжа+ть* можно только что-то делать, чему и отвечает пропозиция). В то же время данные глаголы традиционно трактуются как полнозначные, не отличающиеся от любых других глаголов. При обработке большого числа конструкций с указанными глаголами видно, что они во многом ведут себя как типичные вспомогательные глаголы: как и при использовании вспомогательного глагола *бы+ть*, именно фазовые глаголы выражают грамматическую информацию, а управляемые

ими полнозначные глаголы – лексическую. То же видно на материале безличных конструкций, где также фазовые глаголы передают безличность и другую грамматическую информацию, а управляемый глагол – лексическую, ср. *начина*$^{+}$*ло света*$^{+}$*ть* и т.п. С учётом сказанного словоформа *на*$^{+}$*чало*, например, может получать в корпусе одно из следующих описаний:

на$^{+}$чало{нача$^{+}$ть=VAX,prfc,tran=,act,indc,past,sg,0,n}

на$^{+}$чало{нача$^{+}$ть=VAX,prfc,tran=,act,sbjn,0,sg,0,n}

на$^{+}$чало{нача$^{+}$ть=VAX,prfc,tran=IPS,0,indc,past,0,0,0}

на$^{+}$чало{нача$^{+}$ть=VAX,prfc,tran=IPS,0,sbjn,0,0,0,0}

на$^{+}$чало{нача$^{+}$ть=VV0,prfc,tran=,act,indc,past,sg,0,n}

на$^{+}$чало{нача$^{+}$ть=VV0,prfc,tran=,act,sbjn,0,sg,0,n}

В качестве **основного варианта лексемы** в целом приняты традиционные решения, но в ряде случаев требуется специальный комментарий. Во многом принятие решения по поводу отнесения словоформы к тому или иному частеречному классу обусловливает и решение вопроса об основном варианте лексемы. Так, например, в нашем корпусе субстантивированные прилагательные и причастия, допускающие употребление как в единственном, так и во множественном числе, помещаются в зависимости от формы числа либо в категорию NNS (нарицательные существительные, имеющие форму только единственного числа), либо в категорию NNP (нарицательные существительные, имеющие форму только множественного числа), следовательно, в качестве основного варианта лексемы в каждом случае будет выступать либо форма им.п. ед.ч. либо форма им.п. мн.ч. Например,

больно$^{+}$й{больно$^{+}$й=NNS,m,anim=,sg,nm}

больна$^{+}$я{больна$^{+}$я=NNS,f,anim=,sg,nm}

больны$^{+}$е{больны$^{+}$е=NNP,0,anim=,pl,nm}

В отличие от концепции «Грамматического словаря русского языка» А.А. Зализняка, в представленном морфологическом описании разграничиваются лексемы хле$^{+}$б – хлеба$^{+}$, трава$^{+}$ – тра$^{+}$вы, вино$^{+}$ – ви$^{+}$на и т.п. Отнесённость к грамматическому классу и вопрос об основном варианте лексемы в данном случае решается по аналогии с существительными больно$^{+}$й, больны$^{+}$е.

Глаголы совершенного и несовершенного вида признаются разными лексемами. В лингвистике такая точка зрения представлена. В нашем случае указанное решение сильно облегчает процедуры лемматизации при переходе от словаря словоформ к словарю лексем: различия между глаголами в сов. и несов. в. настолько несистематичны (префиксация, причем разная, и суффиксация, чередования, перегласовка и т.п., имперфективация, с одной стороны, и перфективация – с другой), что формализация всех этих идиосинкразий чрезвычайно затруднительна. Есть и другие резоны в пользу такого решения. Поэтому для словоформ в пределах парадигмы несовершенного вида основным вариантом лексемы

признаётся инфинитив несовершенного вида; для словоформ в пределах парадигмы совершенного вида – соответственно – инфинитив совершенного вида.

Двувидовые глаголы подаются дважды, как омонимы. Например,

ратифици+ровать{ратифици+ровать=VV0,impf,tran=VVI}

ратифици+ровать{ратифици+ровать=VV0,prfc,tran=VVI}

В корпусе последовательно выделяются **идиомы** (под идиомой понимаются неоднословные целостности, или словосочетания, не выводимые по правилам, а потому включаемые в словарь на правах отдельной единицы). В представленной системе категория «идиома» занимает то же место, что и категория «лексема». В качестве отдельной группы в составе широко понимаемых идиом (как целостных единиц, в основе целостности которых – семантика) выделяются и последовательно разграничиваются в корпусе так называемые *составные слова*. Под последними имеются в виду единицы, которые иногда в литературе называют «сочетаниями, эквивалентными слову» (типа *в обни+мку, в голова+х, а_то+* и пр.). Следует заметить, что составной характер таких сочетаний, как *в обни+мку,* носит орфографический характер, с грамматической же точки зрения это слова-наречия. В пределах категории «составные слова» выделяются разрывные и неразрывные (составные) слова (например, *дру+г__о__дру+ге* - с одной стороны, и *в_обни+мку, на_дыбы+, изо_дня+_ в_де+нь* – с другой). Введение категории «составные слова» позволяет снять омонимию на уровне выбора «свободное словосочетание»/«целостная единица» (например, *Дру+г о дру+ге всегда+ позабо+тится* и *Они+ ненави+дели дру+г__дру+га*; *Всё_равно+ о+н не придё+т* и *У него+ получи+лось, что всё+ равно+*), что в конечном итоге позволяет получить более адекватную картину, с одной стороны, связанную с частотностью единиц, с другой стороны – связанную с конкретными морфологическими характеристиками единиц.

По разным причинам все виды **аналитических грамматических форм** в корпусе представлены без объединения компонентов. Так, все глаголы сослагательного наклонения признаются омонимами по отношению к глаголам прошедшего времени. В действительности, конечно, показатель сослагательности – это **одновременно** служебное слово *бы* и форма глагола, **совпадающая** с формой прошедшего времени (т.е. омонимичная ей). Но служебное слово *бы*, как известно, может присоединяться почти к любой словоформе в составе высказывания (составляя с ней единое фонетическое слово). Даже преодолев трудности его автоматического обнаружения, мы должны будем искать в тексте форму на *-л (а/о/и),* т.е. все равно эта форма, совпадающая с формой прош. времени, должна быть помечена как форма сослагательного наклонения.

В случае аналитической формы будущего времени глагол *бы+ть* маркируется как вспомогательный (VAX) с соответствующим морфологическим описанием, а другой компонент, несущий лексическое значение и формально совпадающий с инфинитивом, описывается как обычный инфинитив. Тем самым информация о грамматическом значении глагольной аналитической формы представлена только на уровне вспомогательного глагола.

То же при описании аналитических форм пассива типа *бы+л сде+лан*: форма от *бы+ть* маркируется как вспомогательный глагол, а второй компонент (в данном случае – *сде+лан*) как краткое причастие. По тому же принципу (необъединения компонентов) описываются и аналитические формы сравнительной степени прилагательных, наречий и предикативов: служебный компонент (*са+мый, бо+лее, ме+нее*) описывается как служебное слово (AUX), а второй компонент — как обычное прилагательное, наречие или предикатив.

Выделение категорий **слова, не входящие в современный литературный язык** (UNC), а также **слова, представленные в латинской записи** (NUL), обусловлены

ориентацией на описание современного *литературного* языка, за пределами которого остаётся просторечие, вульгаризмы, жаргонизмы, диалектизмы и пр., а также действием принципа «каждой словоформе – морфологическое описание», в результате чего никакая запись, встречающаяся в тексте, не может быть в этом смысле проигнорирована.

Как можно видеть из представленного материала, при морфологическом описании словоформы конкретные значения грамматических категорий указываются в том случае, если наличествует парадигма в пределах данной категории, в противном случае в позиции соответствующей категории проставляется нуль. Так, нуль проставляется в описании отношения к форме (полная/краткая) у относительных и притяжательных прилагательных, большинства местоимений-прилагательных, не имеющих соотносительных кратких (а значит, не имеющих и полных) форм. По тому же принципу маркируются залоговые значения: если глагол непереходный, тем самым он (и все производные от него формы) находится вне категории залога, и соответственно в морфологическом описании в предусмотренном для категории залога месте проставляется нуль. Прямой связью между категорией переходности и категорией залога обусловлено и решение относить все собственно безличные глаголы типа *тошни*$^{+}$*ть*, *рва*$^{+}$*ть*, *зноби*$^{+}$*ть* к непереходным, тогда как традиционно указанные глаголы рассматриваются как переходные на том основании, что они могут управлять винительным падежом без предлога.

Не является прямым следствием работы с корпусом, но тоже выступает как результат привлечения большого объема материала при необходимости всё «метить» вывод об одушевленности личных местоимений третьего лица. Казалось бы, эти местоимения нейтральны по отношению к признаку «одушевленность/неодушевленность»: *о*$^{+}$*н* можно сказать и о человеке, и о любом предмете (о столе, о самолете). Однако формально эти местоимения все же приходится считать одушевленными: если бы это было не так, мы бы говорили *ви*$^{+}$*жу о*$^{+}$*н (вижу оно*$^{+}$*)*, а не *ви*$^{+}$*жу его*$^{+}$ и т.п. Таким образом, словоформа *его*$^{+}$, например, может иметь в корпусе одно из следующих описаний:

его$^{+}${о$^{+}$н=PNS,3p,m,anim=,sg,ac}
его$^{+}${о$^{+}$н=PNS,3p,m,anim=,sg,gn}
его$^{+}${оно$^{+}$=PNS,3p,n,anim=,sg,ac}
его$^{+}${оно$^{+}$=PNS,3p,n,anim=,sg,gn}

При выборе технологии морфологической разметки текстов мы исходили из того, что данный корпус, подобно словарям или энциклопедиям, не должен иметь ошибок. Поэтому нами принята система разметки с использованием постоянно пополняемого словаря аннотированных словоформ. В этом случае самый первый текст полностью размечается вручную и по нему создается частотный словарь. Вслед за этим полученный словарь дополняется всеми возможными омонимами и используется при разметке последующих текстов с пополнением после каждого следующего текста. При наличии словаря процесс разметки происходит полуавтоматически с помощью специальной программы. Если конкретная словоформа текста представлена в словаре единственным вариантом, ее морфологическое описание переносится в размеченный текст без ведома оператора. При наличии в словаре нескольких омонимов все они предлагаются оператору для выбора. Наконец, отсутствующую в словаре словоформу оператор описывает вручную. Подобный процесс повторяется итеративно для каждого следующего текста, и по мере увеличения объема размеченного корпуса доля чисто ручной разметки сокращается.