А.В. Венцов, Е.В. Грудева

АНАЛИТИЧЕСКИЕ ФОРМЫ В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЛИТЕРАТУРНОГО ЯЗЫКА

Как известно, цели создания большого (национального) аннотированного корпуса текстов предполагают не только разработку надёжного лингвистического инструментария[1], но и создание на основе построенного корпуса новых грамматики и словаря (словарей) данного языка.

В последнее десятилетие наметился особый интерес к этой проблематике и в нашей стране, что, в частности, отразилось в появлении лингвистических корпусов на материале русского языка (**http://www.ruscorpora.ru**; **http://www.narusco.ru**).

Необходимость аннотировать большие массивы текстов поставила перед лингвистами много вопросов как теоретического, так и практического характера. К последней группе вопросов относится, в частности, следующий: как оптимально соотносить такие показатели, как скорость разметки и её качество?

Опыт работы над созданием собственного корпуса русского языка, а также опыт обращения к результатам работы и технологиям, используемым в этой области другими коллективами, показывает, что теоретически более быстрый и в силу этого экономически более выгодный путь использования алгоритми-

[1] Ср.: «Корпусную лингвистику можно считать усовершенство-ванной методикой сбора и обработки материала – традиционного “расписывания” текстов с последующим использованием как-то организованной картотеки для извлечения из “примеров” грамматической, лексикографической и иной информации, для проверки выдвинутых лингвистических гипотез и т.п.» (*Венцов А.В., Грудева Е.В., Касевич В.Б., Ягунова Е.В.* Национальный корпус русского литературного языка: некоторые результаты, приложения и задачи // Научно-техническая информация. Сер. 2. Информационные процессы и системы / Всероссийский ин-т научной и технической информации. М.: ВИНИТИ, 2005. № 6. С. 35).

ческих процедур лингвистического аннотирования не приводит к нужным результатам: доля ошибок в классификации языковых явлений оказывается недопустимо высокой и для достижения нужного результата приходится обращаться к ручной доразметке текстов. Как представляется, к продукции такого рода, как национальный корпус языка, должны применяться такие же требования, как к словарям, энциклопедиям, учебникам и пр., т.е. наличие фактических ошибок здесь недопустимо в принципе. Более трудоёмким, но и более эффективным с точки зрения требуемого результата, является технология полуавтоматической разметки текстов, которая предполагает наличие оператора, разрешающего вопросы, связанные с омонимией в самом широком смысле слова.

Национальный корпус русского литературного языка (далее – НКРЛЯ) создаётся в Лаборатории моделирования речевой деятельности Санкт-Петербургского государственного университета (научный руководитель лаборатории – В.Б. Касевич).

К числу важнейших особенностей НКРЛЯ относятся обязательная акцентуированность всех словоформ в текстах (включая случаи так называемого вторичного ударения типа *лесо^пито+мник*[1], *двацатичетырё^хэта+жный*), систематическое восстановление в правах буквы «ё» и последовательное выделение так называемых составных слов. Под последними понимаются 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 и этимологический характер, с грамматической же точки зрения это слова-наречия. В пределах категории «составные слова» выделяются разрывные и неразрывные (составные) слова

[1] Здесь и далее символ «+» – знак основного ударения, символ ^ – знак вторичного (нефонологического) ударения.

(например, *дру*$^{+}$*г__о__дру*$^{+}$*ге* – с одной стороны, и *в_обни+мку, на_дыбы+, изо_дня+_в_де+нь* – с другой)[1].

Последовательная реализация указанных принципов позволяет снимать омонимию на разных языковых уровнях, что в конечном итоге даёт возможность получить более адекватную картину, связанную, с одной стороны, с частотностью единиц, с другой стороны – с конкретными морфологическими характеристиками единиц[2]. Ср., например:

1) *заступи+тесь*{заступи+ться=VV0,prfc,intr=,0,impr,0,pl,2p,0} (форма императива) и

засту+питесь{заступи+ться=VV0,prfc,intr=,0,indc,futr,pl,2p,0} (форма индикатива, будущего времени), или

2) *на_попа+*{на_попа+=CW1=AV0} (“составное слово”, неразрывное, наречие) и

на{на=PRT},

[1] *Венцов А.В., Грудева Е.В., Касевич В.Б., Ягунова Е.В.* Об идиомах в национальном корпусе русского языка // Международная конференция «Корпусная лингвистика–2004» (12–14 октября 2004 г.). Тезисы докладов. СПб.: Изд-во Санкт-Петербургского университета, 2004; *Венцов А.В., Касевич В.Б., Ягунова Е.В.* Идиома, слово, фонетическое слово // Язык и речь: проблемы и решения: Сб. науч. трудов к юбилею проф. Л.В. Златоустовой / Под ред. Г.Е. Кедровой, В.В. Потапова. М., 2004.

[2] *Венцов А.В., Грудева Е.В., Касевич В.Б.* Морфологическая проблематика в Национальном корпусе русского литературного языка // Международная конференция «Корпусная лингвистика–2004» (12–14 октября 2004 г.). Тезисы докладов. СПб.: Изд-во Санкт-Петербургского университета, 2004; *Венцов А.В., Грудева Е.В., Касевич В.Б., Сведенцова Е.А., Слепокурова Н.А.* О морфологии в Национальном корпусе русского языка // Материалы XXXIII международной филологической конференции (15–20 марта 2004 г., Санкт-Петербург). Вып. 24 «Общее языкознание». Ч. 2. СПб, 2004.

попа+{по+п=NN0,m,anim=,sg,ac} (сочетание предлога *на* и сущ. *поп* в аккузативе).

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 базовый 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.

В отличие от существующей практики (ср., например, корпус ХАНКО) включать в состав так называемых многокомпонентных единиц, кроме «сочетаний, эквивалентных слову», составные числительные и аналитические формы[1], составители НКРЛЯ ограничиваются «составными словами», что позволяет

[1] *Копотев М.* «Несмотря на» «потому что», или многокомпонентные единицы в аннотированном корпусе русских текстов // «Компьютерная лингвистика и интеллектуальные технологии»: Труды международной конференции «Диалог–2004» («Верхневолжский», 2–7 июня 2004 г.) / Под ред. И.М. Кобозевой и др. М., 2004.

последовательно иметь дело с такой лингвистической единицей, как (лексико-грамматическое) слово.

Что же касается аналитических (морфологических) форм в НКРЛЯ, то традиционная аналитическая форма признаётся состоящей из *двух слов*[1]. Поскольку морфологическому описанию в корпусе подлежит каждое лексико-грамматическое слово, то в составе аналитической формы маркируются оба её компонента. Например, в составе аналитической формы *буду читать* «буду» описывается как вспомогательный глагол с соответствующими значениями лица, числа и пр., а «читать» как инфинитив. То же с аналитическими формами компаратива (*более сильный*), суперлатива (*самый большой*), сослагательного наклонения (*взял бы*), пассива (*был сделан*): словам *более*, *самый* приписываются дескрипторы служебного слова, слово *бы* по традиции относится к разряду частиц, слово *был* описывается как вспомогательный глагол, а слова *сильный, большой, взял, сделан* описываются как обычные прилагательные, глагол в сослагательном наклонении и краткое страдательное причастие соответственно. Ср.:

3) *бу+ду чита+ть*

бу+ду{бы+ть=VAX,impf,intr=,0,indc,futr,sg,1p,0}
чита+ть{чита+ть=VV0,impf,tran=VVI,0}

4) *бо+лее си+льный*

бо+лее{бо+лее=AUX=,0,0,0}
си+льный{си+льный=AJ0=,pln,sg,m,nm}

5) *са+мый большо+й*

са+мый{са+мый=AUX=,sg,m,nm}
большо+й{большо+й=AJ0=,pln,sg,m,nm}

[1] *Касевич В.Б.* Элементы общей лингвистики. М., 1977; *Касевич В.Б.* Семантика. Синтаксис. Морфология. М., 1988.

6) *взя+л бы*

взя+л{взя+ть=VV0,prfc,tran=,act,sbjn,0,sg,0,m}
бы{бы=PRT}

7) *бы+л сде+лан*

бы+л{бы+ть=VAX,impf,intr=,0,indc,past,sg,0,m}
сде+лан{сде+лать=VV0,prfc,tran=VVP,psv,sht,0,m,sg,0}

В результате получается, например, что все глаголы сослагательного наклонения признаются омонимами по отношению к глаголам прошедшего времени. В действительности, конечно, показатель сослагательности – это **одновременно** частица *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. всё равно эта форма, совпадающая с формой прош. времени, будет «интересовать» нас как форма сослагательного наклонения.

Таким образом, введя в класс значений соответствующего морфологического дескриптора, содержащего указание на часть речи, такие категории, как служебное слово и вспомогательный глагол, мы получаем возможность извлекать по этим описателям, кроме всего прочего, информацию об аналитических формах.

Немаловажным представляется и тот факт, что при таком подходе к маркированию аналитических (морфологических) форм наблюдается соответствие между теоретическим построением и технологическим решением.