МОРФОЛОГИЯ И СИНТАКСИС В ПРОЕКТЕ "РУССКИЙ СТАНДАРТ" (СОЗДАНИЕ КОРПУСА ГРАММАТИЧЕСКИ РАЗМЕЧЕННЫХ РУССКИХ ТЕКСТОВ)

Б.П. Кобрицов
Российский государственный гуманитарный университет
bobby@ru.ru

В сообщении рассматривается проект по созданию корпуса грамматически размеченных русских текстов. Описываются и сравниваются лингвистические программы, использовавшиеся для осуществления этой задачи на разных этапах. Приводится обоснование того, почему был осуществлен переход от системы, использующей только морфологический анализ, к системе с элементами синтаксиса, а также рассматриваются способы совершенствования программы с целью дальнейшей оптимизации работы.

1.

Необходимость создания корпусов грамматически размеченных текстов для русского языка является вполне очевидной. Для ученых-лингвистов такой корпус в первую очередь предоставляет новые богатые возможности для поиска языковых примеров при проведении лингвистических исследований. Большинство существующих поисковых программ осуществляют поиск либо по подстроке (рассматривая запрос исключительно как набор символов и пытаясь найти точно такую же последовательность в тексте, что, конечно, совершенно исключает нахождение морфологических вариантов слова), либо разбивают слово из запроса на псевдооснову и флексию и в качестве результата выдают все слова, в которые входит такая псевдооснова (таким образом, можно говорить об использовании примитивной морфологии). Однако и такой метод не дает приемлемых результатов. Отдельно стоит упомянуть поисковые системы, использующие морфологический анализ (это поисковая машина "Яндекс", программа Concordance фирмы "Диалинг"). Они выдают хорошие результаты, однако для исследователя намного чаще стоит задача поиска не отдельной словоформы, а списка слов, удовлетворяющих определенным параметрам. Такая задача с успехом решается при использовании грамматически размеченного корпуса текстов, который дает гораздо более широкие возможности для поиска: это и поиск по конкретной словоформе, поиск всех словоформ слова, поиск по набору грамматических признаков и многое другое. Такой размеченный корпус имеет ценность не только для лингвистических исследований, он также может служить и "испытательным полигоном" для различных систем автоматической обработки текста, информационного поиска и т.д.

На сегодняшний день наиболее известным реализованным проектом в области разметки текстов является Британский национальный корпус (BNC), объем которого составляет около 100 млн. словоупотреблений, и где проведены как морфологическая разметка, так и частичный синтаксический разбор[1].

2.

В 2001 году время на базе Центра лингвистической документации (ЦЛД) при Московском независимом университете проводился семинар, посвященный современным проблемам и задачам лингвистики. Тогда родилась идея создания корпуса грамматически размеченных русских текстов - под условным названием "Русский стандарт"(Сичинава, 2003). Первая пробная разметка, на которой проводилась отработка технологии, была

[1] Из российский разработок следует назвать экспериментальные работы по синтаксической разметке текстов, проводимые группой под руководством д.ф.н. И.М.Богуславского в ИППИ РАН (объем обрабатываемых текстов составляет около 100 тыс. слов).

осуществлена при финансовой и технической поддержке компании "Яндекс", объем текстов составил около 3,5 МБ (этот корпус доступен по адресу corpora.yandex.ru). Для "Яндекса" эти размеченные тексты стали полигоном для испытания своей системы информационных запросов, а для "широкой общественности", они стали бы базой для исследования русского языка. Сама разметка осуществляется рабочей группой, состоящей из ученых-лингвистов и студентов лингвистических вузов, под руководством д.ф.н. В.А. Плунгяна.

3.

Итак, первоначально для первичной морфологической разметки была использована программа MyStem, написанная Ильей Сегаловичем, ведущим специалистом компании "Яндекс" по проблемам поиска.

Программа MyStem, как и, наверное, все российские системы морфологического анализа основана на использовании грамматического словаря А.А. Зализняка (Зализняк 1980). Не вдаваясь в тонкости реализации, ее работу можно описать следующим образом: сначала каждая словоформа получает все возможные морфологические интерпретации, которые берутся из базы данных. Однако в таком размеченном тексте существует большое количество неверных разборов, приписанных наиболее часто встречающимся словам (таким как предлоги или частицы). Это связано с тем, что в словаре А.А.Зализняка некоторые морфологически омонимичные слова (формально существующие в русском языке) в реальных текстах никогда, или практически никогда не встречаются. В качестве примера приведем такие частотные случаи как:

- Однобуквенные предлоги (помимо предлогов в словаре они представлены как названия соответствующих букв)

о (нормально без удар.) предл.

о межд.

о с 0 (название буквы о)

же (без удар.) част.; союз

же с 0, (название буквы ж)

- Многозначные существительные, для которых статистически одно из значений встречается на порядок реже, чем другое

ли (без удар.) союз; част

ли с 0 (китайская мера длины)

есть нсв b:

есть предик. (имеется) см. также быть

есть с 0 (старое название буквы е)

есть межд.

- Некоторые формы, формально образованные по номеру парадигмы, однако реально несуществующие или с частотой употребления, стремящейся к нулю

почти (как повелительная форма от *почтить*)

ч'ина (как существительное женского рода)

- А также некоторые другие случаи словоупотреблений, от которых было принято решение отказаться. В частности это случаи неочевидного различия значений (*ведь* част. и *ведь* союз), некоторых случаев субстантивации (*новое* с) и т.д.

Для того, чтобы устранить такие разборы и таким образом снизить объем ручной работы для разметчиков, Алексеем Поляковым был написан специальный постфильтр, задачей которого является отсечение подобных употреблений.

4.

На следующем этапе по работе над проектом "Русский стандарт" планировалось расширить объем корпуса текстов до 10 МБ. Однако как показал опыт ручной разметки начальной части корпуса, такая работа требует значительных

затрат времени. Поэтому мы решили попробовать применить более совершенные механизмы разметки, которые бы оставляли меньше вариантов морфологического разбора.

После рассмотрения различных систем было решено остановиться на программе автоматического анализа текстов, разработанной в компании "Диалинг" под руководством Алексея Сокирко (далее просто Диалинг, см. Сокирко. 2000). При поддержке самих разработчиков система была усовершенствована и дополнена, таким образом, что стала ориентирована на задачу разметки текстов.

Основным отличием программы Диалинг от MyStem является то, что в ней используется автоматический синтаксический анализ. Следует отметить, что основная задача, которую была призвана решить новая система разбора заключалась в том, чтобы снизить объем ручной работы для разметчиков, именно поэтому выбор пал на программу, использующую элементы синтаксического анализа, поскольку заложенные в ней алгоритмы снимают значительную часть омонимии, которую раньше приходилось разрешать человеку.

Как и в случае с MyStem после работы программы разметки Диалинг применяется специальный фильтр. Однако сам фильтр устроен принципиально иначе, так как проблемы, связанные с содержанием словаря, в системе Диалинг можно решить, непосредственно редактируя словарь. Фильтрация требуется в основном исправления стандартных случаев неправильного синтаксического разбора (изменять саму программу синтаксического анализа не представляется возможным).

Рассмотрим теперь некоторые примеры, осуществленной двумя программами, которые показывают преимущества использования синтаксического анализа перед сугубо морфологической разметкой.

Пример 1.

Исходное предложение:

Грязные, горластые крестьяне вели себя совсем не так, как русские мужики.

Разбор MyStem:

Грязные{грязный=A=мн,им|грязный=A=мн,вин,неод}, горластые{горластый=A=мн,им|горластый=A=мн,вин,неод} крестьяне{крестьянин=S,муж,од=мн,им} вели{велеть=V=сов,пов,ед,2-л|велеть=V=несов,пов,ед,2-л|вести=V,несов=прош,мн,изъяв} себя{себя=S,ед,од=род|себя=S,ед,од=вин} совсем{совсем=ADV} не{не=PART} так{так=PART|так=ADV}, как{как=ADV/CONJ}
русские{русский=A=мн,им|русский=A=мн,вин,неод|русская=S,жен,неод=мн,им|русская=S,жен,неод=мн,вин} мужики{мужик=S,муж,од=мн,им}.

Разбор Диалинг:

<st><cl type="Г_ФИН"><gr type="П+С" mw="3"><gr type="ОДН_П" mw="0">*Грязные*{грязный=П=мн,им}, *горластые*{горластый=П=мн,им}</gr> *крестьяне*{крестьянин=С,мр,од=мн,им}</gr> <gr type="ПРЯМ_ДОП" mw="4">*вели*{вести=Г,нс,пе=мн,дст,прш} *себя*{себя=М-С=рд|себя=М-С=вн}</gr> совсем{совсем=Н} не{не=ЧАСТ} так{так=Н|так=СОЮЗ|так=ЧАСТ}</cl> <cl type="0">, как{как=Н|как=СОЮЗ|как=ЧАСТ} <gr type="П+С" mw="12">*русские*{русский=П=мн,им} *мужики*{мужик=С,мр,од=мн,им}</gr> .</cl></st>

В данном предложении представлено несколько видов морфологической омонимии: это омонимия именительного и винительного падежей (прилагательные *грязные, горластые* и *русские*), совпадение форм разных глаголов (*вели* – повелительное наклонение от глагола *велеть* и прошедшая форма от глагола *вести*), омонимия родительного и винительного падежей (*себя*) и омонимия частей речи (*русский*-существительное и *русский*-прилагательное).

Так как в программе MyStem отсутствуют механизмы снятия морфологической неоднозначности, после ее работы над предложением остались все случаи омонимии.

А в разборе Диалинг часть неоднозначности была снята. Так, была убрана омонимичная форма *вели* от глагола *велеть* согласно правилу о том, что в одном предложении не могут находиться глагол в повелительном наклонении и существительное в именительном падеже 3-го лица. Прилагательные *грязные, горластые* были проанализированы как согласованные определения к существительному *крестьяне*, таким образом были уничтожены омонимичные формы винительного падежа, точно так же как и в случае с парой *русские мужики*.

Таким образом в разборе Диалинг осталось лишь восемь морфологических вариантов по сравнению с пятнадцатью у MyStem.

Пример 2.

Исходное предложение:

До вдалеке стоящих кипарисных рощ мы шли двадцать минут.

Разбор MyStem:

До{до=PR} вдалеке{вдалеке=ADV}
стоящих{стоить=V,несов=прич,непрош,мн,род|стоить=V,несов=прич,непрош,мн,вин,од|
стоить=V,несов=прич,непрош,мн,пр}
кипарисных{кипарисный=A=мн,род|кипарисный=A=мн,вин,од|кипарисный=A=мн,пр}
рощ{роща=S,жен,неод=мн,род} мы{мы=S,мн,од=им} шли{слать=V,несов=пов,ед,2-
л|идти=V,несов=прош,мн,изъяв} двадцать{двадцать=NUM=им|двадцать=NUM=вин}
минут{минута=S,жен,неод=мн,род|минуть=V,сов=изъяв,непрош,мн,3-
л|минуть=V,сов=прич,прош,страд,кр,ед,муж}.

Разбор Диалинг:

<st><cl type="Г_ФИН"><gr type="ПГ" mw="0">*До*{до=ПРЕД} <gr type="П+С" mw="4"><gr type="Н+Г"
mw="2">вдалеке{вдалеке=Н}
стоящих{стоить=Г,нс,пе,дст=прч,мн,рд,нст|стоять=Г,нс,нп,дст=прч,мн,рд,нст|стоять=Г,нс,нп,дст=прч,мн,вн,нст|стоять=Г,нс,нп,дст=прч,мн,пр,нст}</gr> *кипарисных*{кипарисный=П=мн,рд} *рощ*{роща=С,жр,но=мн,рд}</gr>
</gr>мы{мы=М-С,мн,1л=им} *шли*{идти=Г,нс,нп,дст=мн,прш} <gr type="Ч+ИГ"
mw="8">двадцать{двадцать=Ч=им|двадцать=Ч=вн} *минут*{минута=С,жр,но=мн,рд}</gr> .</cl></st>

В этом примере имеется морфологическая омонимия между формами прилагательного родительного, предложного и винительного падежа (одушевленного) во множественном числе (*стоящих*[2], *кипарисных*), омонимия глагольных форм (*шли* – повелительное наклонение от глагола *слать* и *шли* – прошедшее время от глагола *идти*), омонимия форм числительного в именительном и винительном падежах (*двадцать*), а также совпадение форм существительного и глагола (*минут* – родительный падеж множественного числа от существительного *минута* и *минут* – страдательное причастие прошедшего времени или форма третьего лица множественного числа настоящего времени от глагола *минуть*). Все эти омонимичные варианты присутствуют в разборе MyStem.

В отличие от этого, в разборе Диалинг были убраны омонимичные формы прилагательного *кипарисный* (так как оно является согласованным определением к существительному *рощ*), убрана форма повелительного наклонения шли от глагола *слать* (см. первый пример), а также оставлен единственный верный разбор словоформы *минут* как формы от существительного *минута* (так как оно входит в группу числительное+ существительное, *двадцать минут*).

В этом примере количество морфологических омонимов у MyStem составило тринадцать, а у Диалинга только шесть (таким образом, так же как и в первом примере разбор Диалинг содержит примерно в два раза меньше морфологических вариантов, чем разбор MyStem).

Как уже говорилось основным достоинство программы разметки Диалинг является использование автоматического синтаксиса. Однако это же и является его основным недостатком. Проблема состоит в том, что варианты словоупотреблений с неснятой омонимией можно отследить, но если синтаксис неверно построил группу, ошибочно удалив остальные варианты, обнаружить такие случаи можно только при сквозном прочтении, что является очень трудоемкой задачей, однако эта процедура необходима на начальном этапе, пока идет обкатка и совершенствование системы, так как позволит в дальнейшем свести количество подобных ошибок к минимуму. Ср. пример неправильного морфологического разбора из-за неверно построенной синтаксической структуры:

Исходное предложение:

Один корнет, правда, читал Ламартина и даже слышал про Шопенгауэра.

Разбор Диалинг:

<st><cl type="0"><gr type="ОДН_ИГ" mw="1"><gr type="П+С" mw="1">Один{один=М-П=мр,ед,им|
один=Ч=мр,им|один=Ч=мр,вн} корнет{корнет=С,мр,но=ед,им|корнет=С,мр,од=ед,им}</gr> ,
правда{правда=С,жр,но=ед,им}</gr> </cl><cl type="Г_ФИН">,<gr type="ПРЯМ_ДОП"
mw="5">читал{читать=Г,нс,пе=мр,ед,дст,прш}
Ламартина{ламартин?=С=мр,фам,ед,рд,од|ламартин?=С=мр,фам,ед,вн,од}</gr> </cl><cl
type="Г_ФИН">и{и=СОЮЗ} даже{даже=ЧАСТ} слышал{слышать=Г,нс,пе=мр,ед,дст,прш} <gr type="ПГ"
mw="10">про{про=ПРЕД}
Шопенгауэра{Шопенгауэра?=С,мр/жр/ср?,од/но?=ед,им|Шопенгауэра?=С,мр/жр/ср?,од/но?=ед,рд|Шопенгауэра?=С

[2] Заметим, что в разборе MyStem для прилагательного *стоящий* отсутствует правильный разбор как причастия от глагола *стоять*. Однако в данном конкретном случае эту пробему можно решить в фильтре.

,мр/жр/ср?,од/но?=ед,дт|Шопенгауэра?=С,мр/жр/ср?,од/но?=ед,вн|Шопенгауэра?=С,мр/жр/ср?,од/но?=ед,тв|Шопенгауэра?=С,мр/жр/ср?,од/но?=ед,пр|Шопенгауэра?=С,мр/жр/ср?,од/но?=мн,им|Шопенгауэра?=С,мр/жр/ср?,од/но?=мн,рд|Шопенгауэра?=С,мр/жр/ср?,од/но?=мн,дт|Шопенгауэра?=С,мр/жр/ср?,од/но?=мн,вн|Шопенгауэра?=С,мр/жр/ср?,од/но?=мн,тв|Шопенгауэра?=С,мр/жр/ср?,од/но?=мн,пр}</gr> .</cl></st>

В данном примере слово "правда" опозналось как член однородного ряда существительных (*корнет, правда*), и все остальные варианты разбора были удалены, хотя правильным разбором является, конечно, вводное слово.

5.

Совершенствовать программу разметки Диалинг можно несколькими путями. Самый очевидный – это дополнение и исправление морфологического словаря, потому что очевидно, что большое количество неверных морфологических вариантов порождает неправильные синтаксические разборы. Например, в системе Диалинг показатель части речи можно приписывать целым сочетаниям слов[3]. Ср.:

<st><cl type="Г_ФИН"><gr type="Н+Г" mw="1">Так{так=Н} прожили{прожить=Г,св,пе=мн,дст,прш}</gr> <gr type="ПРЯМ_ДОП" mw="2">почти{почти=Н} год{год=С,мр,но=ед,им|год=С,мр,но=ед,вн}</gr> </cl><cl type="Г_ФИН">, и{и=СОЮЗ} <gr type="Н+Г" mw="7">славно{славно=Н} прожили{прожить=Г,св,пе=мн,дст,прш}</gr> <cl type="ДПР">, вот{вот=ЧАСТ|вот=ПРДК} уж{уж=Н|уж=С,мр,од=ед,им|уж=ЧАСТ} <gr type="Н+Г" mw="12">воистину{воистину=Н} душа{душить=Г,нс,пе=дпр,дст,нст}</gr> <gr type="ПГ" mw="13">в{в=ПРЕД} душу{душа=С,жр,но=ед,вн}</gr> ,<gr type="ПГ" mw="16">без{без=ПРЕД} <gr type="ОДН_ИГ" mw="17">пошлости{пошлость=С,жр,но=ед,рд} и{и=СОЮЗ} грязи{грязь=С,жр,но=ед,рд}</gr> </gr>.</cl></cl></st>

В данном примере синтаксический анализатор разобрал слово *душа* как деепричастие от глагола *душить*, а если бы в словаре был занесен такой оборот "душа в душу" с грамматической характеристикой *Наречие*, в данном месте ошибки в разборе бы не было.

Кроме того, можно (и эта работа ведется) совершенствовать сам постфильтр. Рассмотрим в качестве примера слова *его*, *ее* и *их*. Они весьма часто попадаются в текстах, однако, после анализа практически всегда остается омонимия между формой притяжательного местоимения (*его дом*) и формой винительного падежа от личного местоимения (*он не узнал ее*). Следующий фильтр, разрешает все случаи такой неоднозначности.

1. Если перед словом[4] стоит предлог, то это слово – притяжательное местоимение (потому что все формы личных местоимений он, она, они при употреблении после предлогов принимают особый префиксальный элемент "н-").

2. Если в предложении находится непереходный глагол, то это слово - притяжательное местоимение.

Пример: *Руки его были скручены за спиной, на шее почему-то висели пустые ножны от шашки, а в углу рта запеклась кровь.*

3. Если в предложении находится переходный глагол, то

если справа от слова стоит существительное (либо группа согласованное прилагательное+существительное)

a. в винительном падеже (или в родительном в случае, если глагол отрицается), **то** это слово - притяжательное местоимение

Пример: *Она увидела его страшные шрамы.*

b. существительное в любом другом падеже, **то** это слово – личное местоимение

Пример: *Варя смерила его презрительным взглядом.*

иначе (справа не существительное)

если в предложении нет других существительных в винительном падеже (или родительном падеже, если глагол отрицается), **то** это слово – личное местоимение

Пример: *Эраст Петрович, у вас под самым носом орудовал опасный, изощренный враг, нанесший нашему делу тяжкий вред и поставивший под угрозу судьбу всей кампании, а вы его так и не распознали.*

[3] Так называемые, обороты – неразрывные фразеологизированные группы слов, которые можно рассматривать как единое целое.

[4] В описании алгоритма под "словом" имеется в виду любое слово из списка {его, ее, их}

иначе – притяжательное местоимение

Пример: *Дей был стар, и женская красота его уже не интересовала…; Знания его больше не приносили пользы.*

Разумеется окончательно избавиться от неправильных разборов скорее всего не удастся, по крайней мере до тех пор, пока не будет написан идеальный синтаксис, однако постоянное совершенствование существующих алгоритмов позволит создать очень мощную систему, которую можно будет использовать не только для морфо-синтаксической разметки текстов, но и для других лингвистических задач.

Литература

1. Зализняк. А.А. Грамматический словарь русского языка, 1980, Москва.
2. Сичинава, Д.В. К проблеме создания корпусов русского языка. 2002
3. Сокирко А.В. диссертационная работа "Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ)"