ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение

высшего образования

Национальный исследовательский университет

«Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа
«Фундаментальная и компьютерная лингвистика»

КУРСОВАЯ РАБОТА

На тему «Автоматическая разметка аналитических форм (буду работать, был сделан, сделал бы, более новый) в НКРЯ.» *Название темы на английском* «Tagging analytic forms in the Russian National Corpus.»

Студентка 2 курса
группы № 141

Сонина Полина Алексеевна

Научный руководитель:
Ляшевская Ольга
Николаевна
Кандидат филологических
наук, профессор Школы
Лингвистики

Москва, 2016

Оглавление

1. Введение

В русском языке форма слова может быть представлена не только одной словоформой, но и сочетанием двух словоформ. Формы, состоящие из знаменательного и служебного слова, называют аналитическими.

В Национальном корпусе русского языка (НКРЯ) морфологическая информация приписывается каждой отдельной словоформе. Следовательно, разметка основного корпуса не отражает то, является ли словоформа частью аналитической формы. Цель данной работы – решить проблему разметки аналитических форм в рамках стандартов НКРЯ.

Задачи нашего исследования состояли в изучении особенностей различных аналитических конструкций; в создании скрипта для автоматической разметки аналитических форм; обнаружении и возможном решении проблем, возникающих при автоматической разметке; уточнение порядка следования элементов конструкции и расстояния между элементами для достижения оптимального баланса между полнотой разметки форм и ее точностью; проверке точности работы созданного скрипта.

2. Аналитические конструкции в русском языке

«Аналитические конструкции (АК) состоят из сочетания основного (полнозначного) и вспомогательного (служебного) слов». «Морфологические АК (аналитические формы) образуют единую словоформу, выражающую морфологическую категорию…». «Семантически и функционально равнозначные слову, они организованы как словосочетания: допускают перестановку компонентов (*Он слушать будет*), включения (*Он будет внимательно слушать*), усечения (*Он будет слушать и записывать*)»[1].

3. Типы аналитических конструкций

Для автоматической разметки были выбраны следующие аналитические конструкции:

1. будущее время глаголов несов. вида (*будет играть*)
2. степени сравнения
    I. сравнительная
        a. прилагательных (*более красивый*)
        b. причастий (*менее подсвечен*)
        c. наречий (*более удачно*)
    II. превосходная
        a. прилагательных (*наиболее яркий*)
        b. причастий (*самый побитый*)
        c. наречий (*быстрее всех*)
3. сослагательное наклонение
    I. инфинитив + *бы/б* (*успеть бы*)
    II. глагол в прош. времени + *бы/б* (*подумали бы*)
    III. предикатив + *бы/б* (*надо бы*)
    IV. *чтобы/чтоб* + инфинитив/гл. в прош.вр. (*чтобы помочь*)
4. страдательный залог (*был устранён*)
5. формы совместного действия
    I. *давай/давайте* + императив (*давайте поможем*)

[1] Лингвистический энциклопедический словарь, Москва, 1990

II. *давай/давайте* + инфинитив (*давай спать*)

В этом списке не представлены все возможные типы аналитических конструкций. Он содержит основные и часто встречающиеся конструкции, которые можно во многих случаях однозначно распознать в тексте с морфологической разметкой.[2] Рассмотрим эти конструкции подробнее.

3.1. Будущее время глаголов несовершенного вида

Аналитическое выражение формы будущего времени свойственно глаголам несовершенного вида. Они образуются при помощи вспомогательного глагола *быть*, стоящего в форме будущего времени, и инфинитива значимого глагола.

Пример употребления (все сентенциальные примеры взяты из НКРЯ):

(1) *Прогреть 3-4 минуты, пока молоко не загустеет и творог не* ***будет тянуться*** *как резина.* [Рецепты национальных кухонь: Эстония (2000-2005)]

В данной работе при создании программы учитываются только некоторые из возможных вариантов порядка словоформ, входящих в состав аналитической формы будущего времени, в предложении: *быть* перед инфинитивом (см. пример (1)); инфинитив перед *быть* (*контролировать буду*); *быть* на расстоянии 1 слова от инфинитива (*будем уверенно действовать*); *быть* перед двумя инфинитивами, соединёнными союзом *и* или *или* (*будут петь и танцевать*).

Подробнее о формах будущего времени в «Русской грамматике»: http://rusgram.narod.ru/1490-1515.html#1493

3.2. Сравнительная степень

Аналитические формы сравнительной степени образуются при помощи прибаления слов *более* и *менее*[3] к простой форме прилагательного (2), причастия (3) или наречия (4).

[2] К аналит формам можно также отнести предложные формы местоимения *друг друга,* ср. *друг от друга.* Разметка этих форм, однако, не входит в нашу задачу, так как их можно также трактовать, как неоднословные лексические единицы (ср. «обороты» в НКРЯ: http://www.ruscorpora.ru/obgrams.html)

[3] В «Русской грамматике» словосочетания с *более* и *менее* не считаются показателями аналитической формы сравнительной степени, так как эти слова сохраняют своё лексическое значение (http://rusgram.narod.ru/1342-1365.html). Но в рамках данной работы будем относить такие сочетания к аналитическим формам.

(2) *Там же, в этих фильмах, воздух* ***более плотный****, цвета такие, или вообще — всё черно-белое и значительное.* [Евгений Гришковец. ОдноврЕмЕнно (2004)]

(3) *Поэтому подъёмная сила шара больше на улице, где воздух* ***менее прогрет****.* [Владимир Лукашик, Елена Иванова. Сборник задач по физике. 7-9 кл. (2003)]

(4) *Кроме того, как заметил Экер, они работают* ***более эффективно****, поскольку топливо не смешивается с водой до поступления в микрореактор.* [Концепция DMFC совершенствуется // «Computerworld», 2004]

В программе учитываются варианты порядка словоформ в предложении, соответствующие отраженным в примерах (2)-(4), то есть более/менее перед значимым словом.

3.3. Превосходная степень сравнения

Аналитические формы превосходной степени образуются путём прибавления слов *наиболее* или *наименее*[4] к простой форме прилагательного (5), причастия (6) или наречия (7); форм слова *самый* к начальной форме прилагательного (8) или причастия (9); словоформы *всех* к сравнительной форме прилагательного (10) или наречия (11).

(5) *Бюджетный процесс — едва ли не центральное звено и уж точно* ***наиболее сложная*** *система в государственном управлении.* [Владислав Кулаков. Уральский САПФИР // «Computerworld», 2004]

(6) *Зато часто они— лишь безумное или нелепое, иногда даже кощунственное сцепление голосов, жестов, мук,* ***наименее гармонирующих*** *друг с другом.* [И. Ф. Анненский. Вторая книга отражений (1909)]

(7) *Во время ночного сна кожа* ***наиболее активно*** *восстанавливается.* [Ворожея. Исполнение желаний // «Даша», 2004]

(8) *Это означает, что система может использоваться в условиях* ***самых жёстких*** *внешних воздействий.* [Наталья Дубова. «Народная» система хранения // «Computerworld», 2004]

[4] Принадлежность слов *наиболее* и *наименее* к показателям аналитической формы превосходной степени опять же спорна, но в данной работе они будут считаться таковыми.

(9) *В искусственном интеллекте* ***самым распространённым*** *на данный момент является определение Т. Грубера [4]:* [С. Г. Керимов. Интеллектуальный поиск информации, основанный на онтологии // «Информационные технологии», 2004]

(10) — *С точки зрения науки, масштабности мышления Энвэ был намного* ***выше всех****.* [Даниил Гранин. Зубр (1987)]

(11) *Все смеются, и родственник без чувства юмора смеётся* ***громче всех****, все довольны.* [Коллекция анекдотов: анекдоты об анекдотах (1970-2000)]

В программе не учитываются варианты порядка словоформ в предложении, отличные от отраженного в примерах (5)-(11).

3.4. Сослагательное наклонение

Сослагательное наклонение, в отличие от остальных представленных в работе категорий, выражается только аналитически. Формы сослагательного наклонения образуются при помощи частицы *бы/б*, которая также может быть в составе союза *чтобы/чтоб*. Частица *бы* чаще всего образует формы сослагательного наклонения в сочетании с инфинитивами (12), с глаголами в прошедшем времени (13), с предикативами (14). В то же время, *бы* может входить в состав эллиптических конструкций с существительными в косвенных падежах и изредка сочетается с причастиями, деепричастиями и императивами, но эти случаи не учитываются в данной работе.

(12) *В Германии часто можно услышать: эх,* ***жить бы*** *в Мюнхене!* [Владимир Гаков. Сказочными дорогами Германии (2001) // «Туризм и образование», 2001.03.15]

(13) *Если бы центральная масса была обусловлена не чёрной дырой, а плотным скоплением обычных звёзд, то ядро галактики* ***светилось бы*** *в десятки раз ярче. [А. М. Черепащук. Поиски чёрных дыр // «Вестник РАН», 2004]*

(14) *В другое время* ***можно бы*** *залюбоваться этой полевой идиллией, но не теперь.* [Василь Быков. Болото (2001)]

В данной работе учитываются только следующие варианты порядка в предложении словоформ, образующих форму сослагательного наклонения: инфинитив, глагол в прошедшем времени или предикатив перед частицей *бы/б*

(см. примеры (12)-(14)); *бы/б* перед инфинитивом или глаголом в прош.вр. (*вот бы уехать, он бы сказал*); союз *чтобы/чтоб* перед инфинитивом или глаголом в прош.вр. (*пришла, чтобы услышать*), а также союз *чтобы/чтоб* на расстоянии одного слова перед инфинитивом или глаголом в прош.вр. (*о том, чтоб они поторопились*).

Если частица *бы/б* стоит перед значимым словом, нужно исключить те случаи, когда инфинитив после *бы/б* является не главным словом, а зависимым от инфинитива, глагола в прошедшем времени или предикатива, стоящего перед частицей и являющегося значимым словом в составе формы сослагательного наклонения. Аналогично, исключаются случаи, когда частица *бы/б* входит в состав союзов *как бы, будто бы, словно бы, точно бы, вроде бы, хотя бы*, так как она не будет являться показателем сослагательного наклонения.

Подобнее о сослагательном наклонении:
http://rusgram.ru/Сослагательное_наклонение

3.5. Страдательный залог

Аналитические формы страдательного залога образуются с помощью формы глагола *быть* и страдательного причастия.

(15) *В феврале* ***был напечатан*** *рассказ "Потенциальный покупатель" Ильи Кочергина, […]* [Алексей Краевский. Журналы и поклонники // «Октябрь», 2003]

В программе учитывается только порядок словоформ, соответствующий порядку в примере 15.

Подробнее о категории залога: http://rusgram.narod.ru/1455-1489.html#1460

3.6. Формы совместного действия

Формы совместного действия являются формами повелительного наклонения, в которых побуждение к действию относится к нескольким лицам, включая говорящего. Аналитические формы совместного действия образуются прибавлением частицы *давай/давайте* к инфинитиву (16) или к глаголу в повелительном наклонении (17).

(16) — *Если вопросов больше нет,* ***давайте пить*** *чай*. [Сергей Довлатов. Наши (1983)]

(17) *Ваши знания **давайте обозначим** вот таким кругом.* [Степан Тимохин (Тим. Собакин). Полное собрание тайн // «Трамвай», 1990]

В работе рассматривается только порядок словоформ, соответствующий примерам 16-17.

Подробнее о формах совместного действия: http://rusgram.narod.ru/1455-1489.html#1479

4. Принципы работы программы

Программа для автоматической разметки аналитических форм написана на языке Python (версия 3.5.1). Используются библиотеки os, re.

4.1. Входные и выходные данные

В качестве материала для разметки был использован корпус текстов со снятой лексико-грамматической омонимией – подкорпус НКРЯ (объём - 533 текста).

На вход программа получает файлы в формате xhtml в кодировке windows-1251, обходя директорию, название которой задано в тексте программы.

При выполнении программы создаётся новая директория, в которой создаются текстовые файлы, в которых была проведена разметка аналитических конструкций. Формат файлов на выходе аналогичен формату входных.

На рисунке 1 представлен образец входных данных. Тексты НКРЯ размечены с помощью следующих тегов:

… - абзацы текста или реплики говорящих;

<se>…</se> - предложения;

<w>…</w> - словоформы;

<ana>…</ana> - разбор словоформы;

параметры тега <ana>: lex – лексема, gr – грамматические признаки.

Словоформа заключена между закрывающим тегами </ana> и </w>.

Ударение в словоформах обозначено символом ` перед ударной гласной.

Знаки препинания стоят после закрывающего тега </w>

В файлах текст заключён между тегами <body>…</body>, а метаинформация внутри <head>…</head>.

…

```
<p><se>
<w><ana lex="мой" gr="A-PRO=m,sg,nom"></ana>Мой</w>
<w><ana lex="брат" gr="S,m,anim=sg,nom"></ana>брат</w>
<w><ana lex="не" gr="PART"></ana>не</w>
<w><ana lex="быть" gr="V,ipf,intr,act=sg,fut,3p,indic"></ana>б`удет</w>
<w><ana lex="играть" gr="V,ipf,tran=inf,act"></ana>игр`ать</w>
<w><ana lex="в" gr="PR"></ana>в</w>
<w><ana lex="фойе" gr="S,n,inan,0=sg,loc"></ana>фой`е</w>!
</se>
…
```

Рис. 1. Фрагмент входного файла.

Представленный фрагмент содержит вхождение аналитической формы будущего времени глагола несовершенного вида – *будет играть*. Задача скрипта опознать форму и добавить тег (*fut_an*) в грамматический разбор глагола *играть*.

4.2. Алгоритм работы программы

При запуске программа запрашивает у пользователя путь к до директории, в которой находятся файлы для разметки, и помещает его в переменную *path*. Затем проверяется наличие заданной папки. В случае отсутствия папки, выводится сообщение об ошибке (*Path does not exist*) и программа завершает работу. Если папка существует, программа продолжает работу.

В той же папке, в которой находится исходная папка с текстами, создаётся новая для сохранения результатов. Она получает название: *«название старой директории» + «_result».*

Программа начинает обход исходной директории. С помощью регулярного выражения определяется название формата файла. Если формат файла - *xhtml,* программа открывает его для дальнейшей работы. Текст из файла помещается в переменную *text1*. Исходный файл закрывается.

Отдельные функции, производящие разметку различных типов конструкций, по очереди применяются к переменной *text1* (Подробнее о функциях в следующем разделе).

Результат работы функций помещается в переменную *text2*.

В созданной для записи результатов директории создаётся файл с названием идентичным названию исходного файла. В файл записывается содержимое переменной *text2*. Файл закрывается.

Цикл повторяется для всех файлов во всех папках исходной директории.

4.3. Функции для разметки отдельных конструкций

В начале программы создаются функции, предназначенные для добавления тега, обозначающего принадлежность значимого слова к определённой аналитической конструкции, в грамматический разбор слова.

Общий принцип работы:

Функция получает на вход текст, который записывается в внутреннюю переменную *text*. С помощью метода *sub*, предоставленного модулем для работы с регулярными выражениями *re*, производится поиск словосочетаний, подходящих под структуру, заданную в регулярном выражении.

В таблице 1 представлены функции программы, структуры аналитических конструкций, распознаваемые регулярными выражениями, и теги для каждой аналитической конструкции. Типы конструкций описаны в разделе 3.

Табл. 1. Функции и теги для разметки аналитических конструкций

Название функции	Аналитическая конструкция	Тег
tag_fut	будущее время глаголов несов. вида (*быть* + инфинитив)	fut_an
tag_fut_2	будущее время глаголов несов. вида (инфинитив + *быть*)	fut_an
tag_fut_3	будущее время глаголов несов. вида (*быть* + любое слово + инфинитив)	fut_an
tag_fut_4	будущее время глаголов несов. вида (*быть* + инфинитив + *и/или* + инфинитив)	fut_an в разбор каждого инфинитива
tag_compar	сравнительная степень (*более/менее* + прилагательное/причастие/наречие)	comp_an
tag_superl	превосходная степень (*наиболее/наименее* + прилагательное/причастие/наречие; *самый* + прил./причастие; прил./причастие/наречие + *всех*)	supr_an

tag_subjunctive	сослагательное наклонение (инфинитив/глагол прош.вр./предикатив + *бы/б*)	subj_an
tag_subjunctive_2	сослагательное наклонение (*бы/б* + инфинитив/глагол прош.вр.)	subj_an
tag_subjunctive_3	сослагательное наклонение (*чтобы/чтоб* + инфинитив/глагол прош.вр.; или: *чтобы/чтоб* + любое слово + инфинитив/глагол прош.вр)	subj_an
tag_passive	страдательный залог (*быть* + страдательное причастие)	pass_an
tag_coop	совместное действие (*давай/давайте* + инфинитив/глагол в указательном наклонении)	coop_an

4.4. Сложные случаи

При разметке аналитических форм нужно учитывать некоторые особенности русского языка. Основные проблемы связаны с так называемым свободным порядком слов. Даже в случае письменного текста порядок словоформ, входящих в состав аналитической конструкции, может быть не стандартным, или же расстояние между ними может сильно варьироваться. В устных текстах, присутствующих в НКРЯ, отклонения могут быть ещё более частыми. В устных текстах также иногда встречаются повторы конструкций или их частей, нестандартно построенные конструкции.

Следующую сложность представляют сочинённые ряды значимых слов в аналитических конструкциях. В данной работе учтён только такой вариант подобных случаев, когда два сочинённых инфинитива входят в состав формы будущего времени (*будут петь и танцевать*).

Теперь рассмотрим некоторые частные проблемы и их решения.

Для разметки сравнительной степени нужно было исключить словосочетания вроде *тем не менее* + прилагательное/причастие/наречие, для этого в регулярное выражение было добавлено дополнительное условие. На рисунке 2 изображено регулярное выражение для разметки форм сравнительной степени. Дополнительное условие выделено полужирным шрифтом.

(?<!<w><ana lex="то" gr="S-PRO,n,sg=ins"></ana>тем</w>\n

<w><ana lex="не" gr="PART"></ana>не</w>\n

) (<w><ana lex="(?:(?:более)|(?:менее))" gr="[0-9A-Za-z=,\-]*?"></ana>[^\s]+?</w>\s?\n\

<w><ana lex=".+?" gr="(?:(?:A=)|(?:[0-9A-Za-z=,\-]*?partcp)|(?:ADV))[0-9A-Za-z=,\-]*?)("></ana>[^\s]+?</w>)

Рис. 2. Регулярное выражение для разметки форм сравнительной степени из функции tag_compar.

При разметке пассивного залога встречаются глаголы на *–бегнуть/-стигнуть/-стынуть* с двоякой формой инфинитива, например, *достигнуть – достичь*. В таких случаях тег добавляется к разбору каждого из инфинитивов (рисунок 3). В текстах могут встречаться и другие случаи неоднозначного разбора словоформ, но они достаточно редки не учтены в данной работе.

…

<w><ana lex="быть" gr="V,ipf,intr,act=sg,fut,3p,indic"></ana>б`удет</w>

<w><ana lex="достигнуть"

gr="V,pf,intr=partcp,pass,praet,brev,m,sg,**pass_an**"></ana><ana lex="достичь"

gr="V,pf,intr=partcp,pass,praet,brev,m,sg,**pass_an**"></ana>дост`игнут</w>

…

Рис. 3. Фрагмент текста с размеченной формой страдательного залога.

Много проблем возникает при разметке форм сослагательного наклонения, некоторые из которых описаны в разделе 3.4. Когда инфинитив после частицы *бы/б* является зависимым словом, или частица *бы/б* входит в состав одного из определённого набора союзов и не является показателем сослагательного наклонения, конструкция размечаться не должна. Для исключения этих случаев в регулярное выражение добавлены необходимые условия, проверяющие идущее перед частицей слово на отсутствие уже добавленного тега и на то, не является ли лексема одной из набора исключающих союзов (рисунок 4).

(?<!хотя)(?<!как)(?<!будто)(?<!словно)(?<!точно)(?<!вроде)(" gr="[0-9A-Za-z=,\-]*?"></ana>[^\s]+?</w>.*?\n

<w><ana lex="бы?" gr="[0-9A-Za-z=,\-]*?"></ana>[^\s]+?</w>\s?\n\

<w><ana lex=".+?" gr="V[0-9A-Za-z=,\-]*?(?:(?:inf)|(?:praet))[0-9A-Za-z=,\-]*?)("></ana>[^\s]+?</w>)

Рис. 4. Регулярное выражение для разметки форм сослагательного наклонения из функции tag_subjunctive_2.

5. Оценка результатов

Для того, чтобы оценить эффективность работы программы, был создан стандарт для оценки качества. Он состоит из 10 файлов, произвольно выбранных из корпуса текстов со снятой лексико-грамматической омонимией, в которых аналитические формы были размечены вручную. Размечены были только те типы конструкций, которые рассматривались в данной работе и для разметки которых предназначена программа.

В таблице 2 представлены результаты работы программы по разметке отдельных типов конструкций: количество тегов, добавленных при ручной разметке; количество верных и лишних тегов, добавленных программой.

Табл. 2. Результаты работы программы на тестовой выборке

Аналитическая конструкция	Количество вхождений (ручная разметка)	Количество правильно распознанных	Количество ошибочно размеченных
будущее время	62	44	0
сравнительная степень	15	10	0
превосходная степень	40	38	1
сослагательное наклонение	222	136	1
страдательный залог	58	49	2
совместное действие	9	4	0

Очевидно, что ни в одном случае программа не смогла разметить все конструкции. Это можно объяснить особенностями русского языка, описанными в разделе 4.4. Количество же ошибочно размеченных конструкций достаточно мало, но нужно учитывать небольшой объём выборки текстов.

В таблице 3 представлены оценки точности (precision) и полноты (recall) автоматической разметки, а также F-мера. В данном случае точность – отношение правильно распознанных конструкций ко всем найденным конструкциям, полнота - отношение правильно распознанных конструкций ко всем вхождениям (размеченным вручную). F-мера объединяет точность P и полноту R по формуле: F=2*P*R/(P+R)

Табл. 3. Оценки точности разметки

Аналитическая конструкция	Точность (precision)	Полнота (recall)	F-мера
будущее время	100,0%	71,0%	83,0%
сравнительная степень	100,0%	66,7%	80,0%
превосходная степень	97,4%	95,0%	96,2%
сослагательное наклонение	99,3%	61,3%	75,8%
страдательный залог	96,1%	84,5%	89,9%
совместное действие	100,0%	44,4%	61,5%
Общая оценка	98,6%	69,2%	81,3%

Оценка точности работы программы велика, так как количество ошибочно размеченных конструкций мало. Полнота разметки показывает, что многие случаи программе разметить не удалось. Всё же показатель полноты высок для наименее гибких конструкций: превосходная степень и страдательный залог. Эти конструкции чаще всего встречаются в предусмотренном в программе виде.

Рассмотрим некоторые случаи ошибочной разметки.

На рисунке 5 представлен фрагмент текста, а котором ошибочно размечена форма превосходной степени. Ошибка возникла из-за того, что словоформа *самую* в данном случае не является показателем превосходной степени, а входит в состав конструкции *ту же самую*. Подобные случаи можно при дальнейшей доработке программы исключить добавлением дополнительного условия в регулярное выражение.

```
…
<w><ana lex="много" gr="NUM=acc"></ana>мн`ого</w>
<w><ana lex="год" gr="S,m,inan=pl,gen"></ana>лет</w>
<w><ana lex="вести" gr="V,ipf,tran=inf,act"></ana>вест`и</w>
<w><ana lex="тот" gr="A-PRO=f,sg,acc"></ana>ту</w>
<w><ana lex="же" gr="PART"></ana>же</w>
<w><ana lex="самый" gr="A-PRO=f,sg,acc"></ana>с`амую</w>
<w><ana lex="просветительский"
gr="A=f,sg,acc,plen,supr_an"></ana>просвет`ительскую</w>
<w><ana lex="работа" gr="S,f,inan=sg,acc"></ana>раб`оту</w> .</se>
…
```

Рис. 5. Фрагмент текста с ошибочной разметкой формы превосходной степени.

В фрагменте на рисунке 6 ошибочно размечена форма сослагательного наклонения. Причастие оказалось размеченным, так как без исключения параметра *partcp* причастие может быть принято программой за форму глагола в прошедшем времени. Подобные ошибки можно исключить добавлением условия в регулярное выражение.

```
…
<w><ana lex="чтобы" gr="CONJ"></ana>чт`обы</w>
<w><ana lex="сломать" gr="V,pf,tran=inf,act"></ana>слом`ать</w>
<w><ana lex="установиться"
gr="V,pf,intr,med=partcp,m,sg,acc,praet,plen,subj_an"></ana>установ`ившийся</w>
<w><ana lex="международный"
gr="A=m,sg,acc,plen"></ana>междунар`одный</w>
<w><ana lex="порядок" gr="S,m,inan=sg,acc"></ana>пор`ядок</w>
…
```

Рис. 6. Фрагмент текста с ошибочной разметкой формы сослагательного наклонения.

В фрагменте на рисунке 7 ошибка при разметке форм страдательного залога возникла из-за того, что глагол *быть* в данном случае не относится к страдательному причастию *согласованные*, а к конструкции *где-то на местах*.

…

<w><ana lex="мочь" gr="V,ipf,intr,act=pl,praes,3p,indic"></ana>М`огут</w>

<w><ana lex="ли" gr="PART"></ana>ли</w>

<w><ana lex="быть" gr="V,ipf,intr,act=inf"></ana>быть</w>

<w><ana lex="согласовать"

gr="V,pf,tran=partcp,pl,nom,pass,praet,plen,**pass_an**"></ana>соглас`ованные</w>

<w><ana lex="список" gr="S,m,inan=pl,nom"></ana>сп`иски</w>

<w><ana lex="где-то" gr="ADV-PRO"></ana>гд`е-то</w>

<w><ana lex="на" gr="PR"></ana>на</w>

<w><ana lex="место" gr="S,n,inan=pl,loc"></ana>мест`ах</w> ?</se>

…

Рис. 7. Фрагмент текста с ошибочной разметкой формы страдательного залога.

В таблице 4 приведены результаты разметки для всего корпуса текстов НКРЯ со снятой лексико-грамматической омонимией (объём - 533 текста), так называемого 1-миллионного золотого стандарта.

Табл. 4. Результаты работы программы на корпусе текстов со снятой лексико-грамматической омонимией

Аналитическая конструкция	Количество вхождений (автоматическая разметка)
будущее время	1119
сравнительная степень	643
превосходная степень	1038
сослагательное наклонение	2496
страдательный залог	1567
совместное действие	58

6. Заключение

В ходе выполнения данной работы были изучены аналитические конструкции в русском языке, их типы и особенности. Был создан скрипт на языке Python для разметки наиболее часто встречающихся конструкций в текстах НКРЯ. Были решены некоторые проблемы, возникающие при разметке.

Дальнейшая разработка скрипта возможна в направлении расширения группы размечаемых конструкций, исключения большего количества ошибочных случаев разметки, увеличения количества учитываемых положений словоформ в предложении.

При достаточной доработке и расширении возможностей, созданный скрипт может быть полезным инструментом при разметке аналитических конструкций. Всё же многие сложные случаи, обусловленные особенностями русского языка (главным образом устной речи), скорее всего, и в дальнейшем потребуют ручной разметки и проверки.

Источники

Лингвистический энциклопедический словарь, Москва, 1990

Русская грамматика (http://rusgram.narod.ru/)

Проект корпусного описания русской грамматики (http://rusgram.ru)

Добрушина Н.Р. Сослагательное наклонение. *Материалы для проекта корпусного описания русской грамматики* (http://rusgram.ru), 2014.

Precision and recall. (2016, May 13). In Wikipedia, The Free Encyclopedia. Retrieved 20:15, May 30, 2016, from https://en.wikipedia.org/w/index.php?title=Precision_and_recall&oldid=720105307

Приложение

Код скрипта на сайте Github:

https://github.com/SoDipole/tagging-analytic-forms