ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 41 за 2013 г.

Клышинский Э.С., Кочеткова Н.А., Мансурова О.Ю., Ягунова Е.В., Максимов В.Ю., Карпик О.В.

# Формирование модели сочетаемости слов русского языка и исследование ее свойств

***Рекомендуемая форма библиографической ссылки:*** Формирование модели сочетаемости слов русского языка и исследование ее свойств / Э.С.Клышинский [и др.] // Препринты ИПМ им. М.В.Келдыша. 2013. № 41. 23 с. URL: http://library.keldysh.ru/preprint.asp?id=2013-41

**О р д е н а  Л е н и н а**
**ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ**
**имени М.В.Келдыша**
**Р о с с и й с к о й  а к а д е м и и  н а у к**

**Э.С.Клышинский, Н.А.Кочеткова, О.Ю.Мансурова,**
**Е.В.Ягунова, В.Ю.Максимов, О.В.Карпик**

# Формирование модели сочетаемости слов русского языка и исследование ее свойств

**Москва — 2013**

**Клышинский Э.С., Кочеткова Н.А., Мансурова О.Ю., Ягунова Е.В., Максимов В.Ю., Карпик О.В.**

Формирование модели сочетаемости слов русского языка и исследование ее свойств

Работа содержит в себе результаты, полученные в ходе извлечения модели сочетаемости слов русского языка. Приведен метод формирования словаря сочетаемости и выделения из него списка управления для глаголов. Так как основой модели является использование неомонимичных фрагментов текста, выделенных с использованием поверхностного синтаксического анализатора, проведено исследование свойств омонимии в текстах на русском и английском языках. Выделенная база синтаксически сочетающихся слов применена для решения задачи определения стиля текста.

***Ключевые слова:*** сочетаемость слов, модель управления глагола, определение стиля текста

**Klyshinsky E.S., Kochetkova N.A., Mansurova O.Yu., Iagounova E.V., Maximov V.Yu., Karpik O.V.**

Development of Russian subcategorization frames and its properties investigation

The paper describes results collected during investigations of Russian subcategorization frame. Methods for subcategorization frame and for verbal government extraction are also proposed in the paper. Investigations of homonymy in Russian and English languages were provided because of the proposed method used unambiguous words only. Collected database of syntactic dependent words combinations is applied to the task of texts' style detection.

The research is partially supported by RFH grant №12-04-00060-a, RFBR grant № 11-01-00793.

***Key words:*** words co-occurrence, verbal government model, text style detection

Работа выполнена при поддержке РГНФ, грант № 12-04-00060-а, РФФИ, грант № 11-01-00793-а

## Введение

Глагольное управление является важной частью систем автоматической обработки текстов. Информация о глагольном управлении используется для разрешения омонимии [Толдова 2008, Клышинский 2011], снятия синтаксической неоднозначности [Гельбух 1999] и синтаксического анализа в целом [Волкова 2003]. В данной статье нас больше будут интересовать вопросы создания словаря глагольного управления.

Под моделью управления (МУ) глагола в данной статье будем понимать информацию о сочетаемостных характеристиках глагола. То есть глагольное управление показывает, какими грамматическими (род, число, падеж) или семантическими (связи слова в тезаурусе или онтологии) параметрами должно обладать существительное, для того чтобы быть связанным при помощи подчинительной связи с данным глаголом через заданный предлог или без него. В данной работе мы обрабатываем только лексическую информацию, привязанную к словам, а конкретнее – падеж существительных.

Таким образом, в нашем случае словарь МУ будет содержать в себе информацию о том, в каком падеже может находиться существительное, присоединяемое к данному глаголу через заданный предлог. Заметим, что подобное определение не делает разницы между актантами и сирконстантами.

Под словарем сочетаемости будем понимать информацию о том, с какими словами может быть синтаксически связано данное слово в различных текстах.

## Существующие решения

На данный момент в области получения словарей глагольного управления работы ведутся в двух направлениях: автоматическое и ручное составление словарей. Составленные вручную словари содержат в себе информацию высокого качества, но обладают малым объемом. Так, например, работа [Денисова 2002] содержит в себе всего 2500 статей, хотя и весьма представительных, приводящих не только информацию о сочетании слова с другими, но и толкования данного слова, его грамматические характеристики. Объем более ранних печатных работ, специализированных на глагольном управлении [Розенталь 1986, Апресян 1982] также не очень велик.

Для решения указанной задачи возможно привлечение имеющихся электронных словарей. Словарь «КроссЛексика» [Большаков 2011] содержит в себе около 2 млн. связей, для каждой из которых приведены примеры. Однако для решения задачи создания словаря моделей глагольного управления данный ресурс не применялся. Более скромный, но более доступный словарь [Бирюк 2012] содержит 10 000 статей, но всё еще не охватывает значительную часть современной лексики.

Часть приведенных выше словарей, вместе с информацией, доступной в Национальном корпусе русского языка (НКРЯ), использовались для

составления электронных словарей [Кобрицов 2007]. Развитием данного метода стал ресурс Фреймбанк [Framebank 2012]. Данный ресурс хранит информацию не только о модели управления, но и о типе синтаксической связи – порядка 27 000 пар [Ляшевская 2009]. Таким образом, в проекте Фреймбанк хранится меньше информации, хотя она находится на качественно ином уровне.

Автоматическое извлечение словаря МУ уже предлагалось в ряде работ как для русского [Гельбух 1999], так и для других языков [Manning 1993, Messiant 2008, Preiss 2007]. В них предлагалось использовать конечные автоматы (КА) для распознавания отдельных цепочек «глагол + группа существительного» [Manning 1993] или системы синтаксического анализа [Гельбух 1999, Messiant 2008, Preiss 2007]. Далее предлагалась процедура, выделяющая из полученных результатов синтаксические связи между глаголом (а в случае [Preiss 2007] – и существительным) и зависимыми словами.

Однако относительно невысокое качество синтаксического анализа и морфологическая неоднозначность приводили к тому, что результат включал значительное число ошибок (от 5 до 20%), требующее ручного вмешательства. Большое количество выделяемых зависимостей делало создание словаря МУ практически нереализуемым, малое количество связей не давало выигрыша по сравнению с бумажными словарями.

Для устранения этой проблемы могут использоваться размеченные корпуса. На данный момент НКРЯ содержит 17,5 млн. предложений, содержащих почти 210 млн. словоупотреблений [НКРЯ 2012]. Омонимия снята лишь с 516 000 предложений, содержащих почти 6 млн. словоупотреблений. Самый крупный на данный момент синтаксически размеченный корпус СинТагРус по состоянию на 2011 год содержал более 45 000 синтаксически размеченных предложений в середине года [Frolova 2011] и более 49 000 предложений на конец года [СинТагРус 2011]. Заметим, что по нашим оценкам для составления словаря МУ для 25 000 – 30 000 глаголов требуется корпус примерно в 6 миллионов предложений, то есть объемов существующих корпусов всё еще недостаточно.

В данной работе предлагается подход, использующий неразмеченный корпус текстов большого объема и позволяющий повысить качество извлекаемых МУ. Подход основывается на применении КА и использовании неомонимичной части корпуса.

## Выделяемые конструкции

Для устранения синтаксической неоднозначности взамен построения дерева зависимостей для всего предложения было решено использовать только группы, синтаксическая корректность которых не вызывает сомнения в подавляющем большинстве случаев. В качестве таких групп брались, например, следующие (в качестве глагола могут использоваться деепричастия, а при связи через предлог – и причастия):

1. Группа существительного, следующая за единственным глаголом в предложении, синтаксически подчиняется данному глаголу, при этом прилагательные подчиняются существительному.

2. Единственная группа существительного, расположенная в начале предложения перед единственным глаголом, синтаксически подчиняется данному глаголу, при этом прилагательные подчиняются существительному.

3. Прилагательные, расположенные между предлогом и существительным, синтаксически подчиняются данному существительному.

4. Если после существительного, находящегося не в родительном падеже, следует существительное в родительном падеже, причем между первым существительным и глаголом находится предлог, то второе существительное подчиняется первому.

5. В конструкции «глагол + предлог + существительное/глагол» третье слово можно считать глаголом. Тогда существительное подчиняется глаголу через предлог.

6. Группа слов вида «предлог + прилагательное + прилагательное/существительное + существительное», в которой согласование возможно только при условии, что второе слово считается прилагательным, считается группой существительного, и все прилагательные подчиняются последнему слову.

## Метод сбора информации о глагольном управлении

В предлагаемом подходе выделение конструкций ведется с использованием КА. По данным правилам получается значительно больший объем информации, чем это необходимо для составления словаря МУ глаголов.

*Шаг 1* – извлечение сочетаний из текста. На данном шаге проводится морфологическая разметка корпуса текстов. Далее по шаблонам 1 и 2 отбираются глагольные сочетания. В формировании данной базы принимают участие существительные, однозначные по части речи, но, возможно, неоднозначные по падежу. В связи с этим полученные сочетания в чистом виде еще не могут использоваться для создания словаря глагольных МУ.

Проиллюстрируем предложенный метод на примере. Итак, пусть имеется некоторый текст, из которого могут быть извлечены, среди прочего, следующие связанные сочетания вида «глагол (+предлог) +существительное».

```
состоятся вечера
приглашает на концерт
исполнят произведения
состоится встреча
примут участие
откроется выставка
```

*Шаг 2* – составление базы сочетаемости слов. Полученные на предыдущем шаге сочетания приводятся к нормальной форме, после чего рассчитывается их

встречаемость. Из полученной базы отсеиваются сочетания, встречаемость которых ниже заданного порога. В нашем случае мы получим следующую базу.

```
ПРИГЛАШАТЬ;НА;КОНФЕРЕНЦИЯ;218
ПРИГЛАШАТЬ;НА;КОНЦЕРТ;281
ПРИГЛАШАТЬ;НА;КОНЬЯК;3
ПРИГЛАШАТЬ;НА;КОРАБЛЬ;17
ПРИГЛАШАТЬ;НА;КОРДОН;3
ПРИГЛАШАТЬ;НА;КОРОНАЦИЯ;6
```

Здесь показаны абсолютные значения встречаемости указанных сочетаний, приведенных к нормальной форме.

## Исследование омонимии в русском языке

Как это отмечалось выше, в отличие от предыдущих работ, мы используем только группы неомонимичных слов, составляющих синтаксически связанные конструкции. В связи с этим необходимо понять, какую вообще часть текстов мы имеем возможность проанализировать. Для этого были проведены эксперименты по сбору статистики для различных видов омонимии. Исследовался процент омонимичных и неомонимичных слов (остальные слова не были распознаны системой морфологического анализа). Для омонимичных слов исследовались следующие параметры: какой процент слов омонимичен по нормальной форме, по части речи, по обоим видам или для одного и того же слова неоднозначен набор параметров. Полученные результаты представлены в Таблице 1. Как видно из таблицы, различные корпуса обладают сходным распределением, хотя и зависящим от жанра и стиля текста. Так, например, научные тексты, судя по имеющимся данным, обладают большей омонимией, однако набирается она за счет использования слов, имеющих одинаковое написание при разных наборах параметров. Кроме того, в художественных текстах авторы меньше внимания уделяют устранению частеречной омонимии.

Мы решили также проверить гипотезу о том, что распределение слов может колебаться в зависимости от объема выборки внутри одного корпуса или периода выборки. В Таблице 2 представлены результаты обработки корпуса статей газеты «Коммерсант» за различные периоды. Как видно из таблицы, данное предположение не подтверждается (по крайней мере, на современных материалах). Как видно из таблиц, в зависимости от корпуса в текстах имеется 35 – 50% неомонимичных словоупотреблений, то есть вероятность найти описанные конструкции достаточно велика.

Заметим, что омонимия в русском языке существенно отличается от омонимии английских текстов (см. Табл. 3). Очевидно, что проблема снятия омонимии в английском языке стоит более остро, а сама омонимия носит качественно иной характер. Если в русском языке основной проблемой является разрешение значений параметра одного слова, то в английском языке важным является определение части речи слова.

*Таблица 1*

**Распределение видов неоднозначности в русском языке для разных корпусов**

| Название | Объем, млн токенов | Однозн. | Неоднозначные | | | | |
|---|---|---|---|---|---|---|---|
| | | | Всего | По части речи (чр) | По нач. форме (нф) | По чр и нф | По пар. |
| РИА Новости 2001 – 2009 | 154,2 | 47,2% | 48,7% | 5,4% | 4,7% | 9,4% | 29,2% |
| Лента 2005 – 2009 | 32,4 | 48,2% | 47,4% | 5,3% | 4,7% | 10% | 27,7% |
| Компьюлента 2001 – 2009 | 9,2 | 46,6% | 49,1% | 6,3% | 4% | 10,6% | 28,2% |
| PCWeek/RE 2000 – 2009 | 27,9 | 44,2% | 51,8% | 7,4% | 4,1% | 11,4% | 28,9% |
| Библиотека Мошкова | 560,3 | 46,8% | 48,6% | 9,8% | 4,9% | 14,3% | 19,6% |
| Корпус беллетристики | 93,8 | 49% | 47,2% | 9,5% | 4,9% | 14,2% | 18,6% |
| Корпус диссертаций 1 | 12,5 | 37,2% | 56,8% | 7,5% | 4,6% | 9,3% | 35,4% |
| Корпус диссертаций 2 | 11,2 | 37,4% | 54,2% | 6,9% | 4,6% | 8,1% | 34,6% |

*Таблица 2*

**Распределение видов неоднозначности в русском языке для разных корпусов**

| Название | Объем, млн токенов | Однозн. | Неоднозначные | | | | |
|---|---|---|---|---|---|---|---|
| | | | Всего | По чр | По нф | По чр и нф | По пар. |
| Коммерсант 07.2007–07.2008 | 23,8 | 46,2% | 48,8% | 7,2% | 4,8% | 11,5% | 25,3% |
| Коммерсант 08.2008 | 1,7 | 46% | 48,8% | 7,1% | 4,8% | 11,2% | 25,7% |
| Коммерсант 09.2008–04.2009 | 12,4 | 46,1% | 48,8% | 7,2% | 4,8% | 11,5% | 25,3% |
| Коммерсант 09.2008–07.2009 | 17,5 | 46,1% | 48,8% | 7,2% | 4,8% | 11,5% | 25,3% |
| Коммерсант 01.2011–04.2012 | 30,7 | 46,1% | 48,7% | 7,1% | 4,8% | 11,3% | 25,5% |

*Таблица 3*

**Сравнение видов неоднозначности для русского и английского языков**

| Название | Объем, млн токенов | Однозн. | Неоднозначные | | | | |
|---|---|---|---|---|---|---|---|
| | | | Всего | По части речи (чр) | По нач. форме (нф) | По чр и нф | По пар. |
| РИА Новости 2001 – 2009 | 154,2 | 47,2% | 48,7% | 5,4% | 4,7% | 9,4% | 29,2% |
| Reuters 2009 | ~300 | 38,9% | 53,5% | 50,4% | 0,32% | 1% | 2,8% |

*Таблица 4*

**Распределение видов неоднозначности в русском языке для корпуса РИА Новости**

| | noun | adj | participle | deepr | verb | pers. pronoun | poss. pronoun | dem. pronoun | cardinal number | ordinal number | adv | predic | prep | conj | interj | particle |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| noun | 0 | 0,6455 | 0,1141 | 0,0127 | 0,1588 | 0,0063 | 0,2151 | 0,0034 | 0,0084 | 0,0980 | 0,0501 | 0,0174 | 0,0312 | 0,0208 | 0,0003 | 0,0091 |
| adj | 0,6455 | 0 | 0,4123 | 0,0002 | 0,0032 | 0,0019 | 0,0736 | 0 | 0,0016 | 0 | 0,0628 | 0,0174 | 0,0052 | 0,0024 | 0 | 0,0122 |
| participle | 0,1141 | 0,4123 | 0 | 0 | 0,0012 | 0 | 0,0765 | 0 | 0 | 0 | 0,0004 | 0 | 0 | 0 | 0 | 0 |
| deepr | 0,0127 | 0,0002 | 0 | 0 | 0 | 0 | 0,0007 | 0 | 0 | 0 | 0,0001 | 0 | 0,0266 | 0,0013 | 0 | 0,0013 |
| verb | 0,1588 | 0,0032 | 0,0012 | 0 | 0 | 0 | 0,0114 | 0 | 0,0096 | 0 | 0,0045 | 0,0289 | 0,0123 | 0 | 0,0289 | 0,0001 |
| pers. pronoun | 0,0063 | 0,0019 | 0 | 0 | 0 | 0 | 0,1182 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| poss. pronoun | 0,2151 | 0,0736 | 0,0765 | 0,0007 | 0,0114 | 0,1182 | 0 | 0,2892 | 0,0096 | 0 | 0,0137 | 0,0243 | 0 | 0,0322 | 0 | 0,0617 |
| dem. pronoun | 0,0034 | 0 | 0 | 0 | 0 | 0 | 0,2892 | 0 | 0,0094 | 0 | 0,0072 | 0,0009 | 0 | 0,0073 | 0 | 0,0046 |
| cardinal number | 0,0084 | 0,0016 | 0 | 0 | 0,0096 | 0 | 0,0096 | 0,0094 | 0 | 0 | 0,0132 | 0,0031 | 0 | 0 | 0 | 0 |
| ordinal number | 0,0980 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| adv | 0,0501 | 0,0628 | 0,0004 | 0,0001 | 0,0045 | 0 | 0,0137 | 0,0072 | 0,0132 | 0 | 0 | 0,0150 | 0,0310 | 0,0224 | 0 | 0,0308 |
| predic | 0,0174 | 0,0174 | 0 | 0 | 0,0289 | 0 | 0,0243 | 0,0009 | 0,0031 | 0 | 0,0150 | 0 | 0,0057 | 0,0034 | 0,0042 | 0,0055 |
| prep | 0,0312 | 0,0052 | 0 | 0,0266 | 0,0123 | 0 | 0 | 0 | 0 | 0 | 0,0310 | 0,0057 | 0 | 0 | 0 | 0 |
| conj | 0,0208 | 0,0024 | 0 | 0,0013 | 0 | 0 | 0,0322 | 0,0073 | 0 | 0 | 0,0224 | 0,0034 | 0 | 0 | 0 | 0 |
| interj | 0,0003 | 0 | 0 | 0 | 0,0289 | 0 | 0 | 0 | 0 | 0 | 0 | 0,0042 | 0 | 0 | 0 | 0 |
| particle | 0,0091 | 0,0122 | 0 | 0,0013 | 0,0001 | 0 | 0,0617 | 0,0046 | 0 | 0 | 0,0308 | 0,0055 | 0 | 0 | 0 | 0 |

adj – прилагательное
adv – наречие
cardinal number – количественное числительное
conj – союз
deepr – деепричастие – форма глагола
dem. pronoun – указательное местоимение
interj – междометье
noun – существительное
ordinal number – порядковое числительное
participle – причастие – форма глагола
particle – частица
pers. pronoun – личное местоимение
poss. pronoun – притяжательное местоимение
predic – предикатив
prep – предлог
verb – глагол

*Таблица 5*

**Распределение видов неоднозначности в русском языке для библиотеки Мошкова**

| | noun | adj | participle | deepr | verb | pers. pronoun | poss. pronoun | dem. pronoun | cardinal number | ordinal number | adv | predic | prep | conj | interj | particle |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| noun | 0 | 0,3441 | 0,023 | 0,0065 | 0,1326 | 0,0552 | 0,1457 | 0,0044 | 0,0044 | 0,0487 | 0,0665 | 0,0227 | 0,0221 | 0,0228 | 0,0014 | 0,0125 |
| adj | 0,3441 | 0 | 0,1282 | 0,0002 | 0,0057 | 0,0032 | 0,0046 | 0 | 0,0038 | 0 | 0,0994 | 0,0281 | 0,0019 | 0,0035 | 0,0003 | 0,02 |
| participle | 0,0229 | 0,1282 | 0 | 0 | 0,0005 | 0 | 0,0064 | 0,0002 | 0 | 0 | 0,0007 | 0 | 0 | 0 | 0 | 0 |
| deepr | 0,0065 | 0,0002 | 0 | 0 | 0 | 0 | 0,0014 | 0 | 0 | 0 | 0,0014 | 0 | 0,0111 | 0,0027 | 0 | 0,0027 |
| verb | 0,1326 | 0,0057 | 0,0005 | 0 | 0 | 0 | 0,0177 | 0,0002 | 0,0054 | 0 | 0,0067 | 0,033 | 0,0042 | 0,0003 | 0,0326 | 0,0006 |
| pers. pronoun | 0,0552 | 0,0031 | 0 | 0 | 0 | 0 | 0,1498 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| poss. pronoun | 0,1456 | 0,0046 | 0,0064 | 0,0014 | 0,0177 | 0,1498 | 0 | 0,4269 | 0,0049 | 0 | 0,0086 | 0,0148 | 0 | 0,0329 | 0 | 0,0788 |
| dem. pronoun | 0,0044 | 0 | 0,0002 | 0 | 0,0002 | 0 | 0,4269 | 0 | 0,0047 | 0 | 0,0138 | 0,0042 | 0 | 0,012 | 0 | 0,003 |
| cardinal number | 0,0044 | 0,0039 | 0 | 0 | 0,0054 | 0 | 0,0049 | 0,0047 | 0 | 0 | 0,0193 | 0,0057 | 0 | 0 | 0 | 0 |
| ordinal number | 0,0487 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| adv | 0,0665 | 0,0994 | 0,0007 | 0,0014 | 0,0067 | 0 | 0,0086 | 0,0138 | 0,0193 | 0 | 0 | 0,0339 | 0,0174 | 0,0441 | 0,0002 | 0,0517 |
| predic | 0,0227 | 0,0281 | 0 | 0 | 0,033 | 0 | 0,0148 | 0,0042 | 0,0057 | 0 | 0,0339 | 0 | 0,0045 | 0,0068 | 0,0051 | 0,0129 |
| prep | 0,0221 | 0,0019 | 0 | 0,011 | 0,0042 | 0 | 0 | 0 | 0 | 0 | 0,0175 | 0,0045 | 0 | 0 | 0 | 0 |
| conj | 0,0228 | 0,0035 | 0 | 0,0027 | 0,0003 | 0 | 0,0329 | 0,012 | 0 | 0 | 0,0441 | 0,0068 | 0 | 0 | 0 | 0 |
| interj | 0,0014 | 0,0003 | 0 | 0 | 0,0326 | 0 | 0 | 0 | 0 | 0 | 0,0002 | 0,0051 | 0 | 0 | 0 | 0 |
| particle | 0,0125 | 0,02 | 0 | 0,0027 | 0,0006 | 0 | 0,0788 | 0,00297 | 0 | 0 | 0,0517 | 0,0128 | 0 | 0 | 0 | 0 |

Для разметки русской части использовалась система морфологического анализа «Кросслятор» [Елкин 2003], для разметки английской части – библиотека Pymorphy [Pymorphy 2013], основанная на словарях с сайта aot.ru. Заметим, что также нами проводилась разметка других текстов с использованием морфологического модуля для русского языка, взятого с сайта АОТ, которая подтвердила полученные закономерности, хотя и показала зависимость результатов от применения конкретного словаря. То есть в целом конкретные цифры будут отличаться для корпусов различной тематики и стилистики, используемых систем морфологического анализа, но приведенные цифры в целом полно отражают разницу между омонимией в русском и английском языках.

Детальное изучение данных показало, что распределение омонимичных слов по типам также зависит от жанра корпуса. Так, в Таблицах 4 и 5 приведены доли видов омонимии для некоторых частей речи. В них показаны только те слова, в которых наблюдалась омонимия между различными частями речи. Расчет данных проводился следующим образом. По результатам морфологического анализа получается набор словоформ, каждой из которых приписана некоторая часть речи. Будем считать, что набор словоформ упорядочен. Если количество частей речи в данном наборе больше одной, то для каждой словоформы берется другая словоформа с отличающейся частью речи и находящаяся дальше в списке. Для этой пары частей речи находится ячейка в таблице, в которую добавляется единица. По окончании морфологического анализа была проведена нормировка матрицы на общее количество словоупотреблений, которые омонимичны по части речи. То есть абсолютные цифры в матрице могут интерпретироваться как количество пар словоформ, омонимичных друг другу. Относительные цифры могут быть проинтерпретированы как доля омонимичных пар для всех случаев омонимии. Несмотря на неудачность меры, она позволяет оценивать вклад данного вида омонимии в общую картину неоднозначности.

В итоге можно заметить, что наибольший вклад вносит омонимия существительное / прилагательное. Далее идут прилагательное / причастие, указательное местоимение / притяжательное местоимение, существительное / притяжательное местоимение, другие виды омонимии. Часть из этих видов омонимии отражает конкретные особенности выбранного морфологического словаря. Кроме того, наблюдения показали, что распределение омонимии по видам также зависит от стиля текста. Приведенные в Таблицах 4 и 5 данные являются достаточно характерными.

Также были проанализированы различные случаи омонимии внутри одной части речи по какому-либо параметру. Результаты, полученные в ходе экспериментов, приведены в Табл. 6 – 8. Полученные результаты позволяют осознанно подойти к проблеме снятия омонимии в русском языке. Зная наиболее вероятные варианты омонимии, можно написать набор правил, снимающих омонимию с большего процента случаев.

*Таблица 6*

**Доля случаев омонимии по падежам внутри одного существительного (РБК)**

| | eso depende | им | род | дат | вин | тв | предл |
|---|---|---|---|---|---|---|---|
| eso depende | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| им | 0 | 0 | 0,14013 | 0,04503 | 0,28311 | 0,00011 | 0,04635 |
| род | 0 | 0,14013 | 0 | 0,06595 | 0,17423 | 0,00025 | 0,06551 |
| дат | 0 | 0,04503 | 0,06595 | 0 | 0,04497 | 0,00024 | 0,08750 |
| вин | 0 | 0,28311 | 0,17423 | 0,04497 | 0 | 0,00012 | 0,04627 |
| тв | 0 | 0,00011 | 0,00025 | 0,00024 | 0,00012 | 0 | 0,00023 |
| предл | 0 | 0,04635 | 0,06551 | 0,08750 | 0,04627 | 0,00023 | 0 |

*Таблица 7*

**Доля случаев омонимии по падежам внутри одного прилагательного (РБК)**

| | eso depende | им | род | дат | вин | тв | предл |
|---|---|---|---|---|---|---|---|
| eso depende | 0 | 0,00019 | 0,00015 | 0,00015 | 0,00019 | 0,00015 | 0,00015 |
| им | 0,00019 | 0 | 0,00582 | 0,00577 | 0,13667 | 0,00577 | 0,00577 |
| род | 0,00015 | 0,00582 | 0 | 0,08190 | 0,14770 | 0,08264 | 0,16190 |
| дат | 0,00015 | 0,00577 | 0,08191 | 0 | 0,00578 | 0,10804 | 0,08188 |
| вин | 0,00019 | 0,13667 | 0,14770 | 0,00578 | 0 | 0,00577 | 0,08171 |
| тв | 0,00015 | 0,00577 | 0,08264 | 0,10804 | 0,00577 | 0 | 0,08188 |
| предл | 0,00015 | 0,00577 | 0,16190 | 0,08188 | 0,08171 | 0,08188 | 0 |

*Таблица 8*

**Доля случаев омонимии по падежам внутри одного причастия (РБК)**

| | eso depende | им | род | дат | вин | тв | предл |
|---|---|---|---|---|---|---|---|
| eso depende | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| им | 0 | 0 | 0 | 0 | 0,17942 | 0 | 0 |
| род | 0 | 0 | 0 | 0,05760 | 0,18707 | 0,05760 | 0,18541 |
| дат | 0 | 0 | 0,05760 | 0 | 0 | 0,09183 | 0,05760 |
| вин | 0 | 0,17942 | 0,18707 | 0 | 0 | 0 | 0,12587 |
| тв | 0 | 0 | 0,05760 | 0,09183 | 0 | 0 | 0,05760 |
| предл | 0 | 0 | 0,18541 | 0,05760 | 0,12587 | 0,05760 | 0 |

## Стилистические особенности текстов

Тексты разных функциональных стилей отличаются по частотности синтаксических конструкций [Герд 1996]. На данном этапе единицей анализа является коллекция в целом. Были взяты коллекции текстов трех функциональных стилей: художественных, новостных и научных. Объем каждой из коллекций от 0,7 млн до 6 млрд словоупотреблений. Объем основных коллекций приведен в табл. 9.

*Таблица 9*

**Коллекции, использованные в экспериментах по определению стиля текста**

| Источник | Объем (млн. словоупотреблений) |
|---|---|
| Художественные тексты | |
| Библиотека Мошкова | 688 |
| WebReadings | 3000 |
| Librusec | 6000 |
| Новостные источники | |
| Лента, 2005–2010 гг. | 42,5 |
| РИА Новости | 186,8 |
| РосБизнесКонсалтинг, 2003–2010 гг. | 28,8 |
| Независимая Газета, 1999–2010 гг. | 100,6 |
| Российская Газета, 2000–2010 гг. | 42 |
| Взгляд | 72,4 |
| Мембрана | 3 |
| Открытые системы | 3,6 |
| Смешанные стили и жанры | |
| ItHappens | 0,7 |
| Популярная механика | 2,1 |
| Научные тексты различных стилей и предметных областей | |
| Журнал «Информатика и системы управления», 2001–2010 гг. | 1 |
| Журнал «Программные продукты и системы», 1998–2010 гг. | 2,5 |
| Коллекция авторефератов различных областей науки | 31,4 |
| Конференция КИИ | 1,1 |
| Конференция RCDL | 1,3 |

Анализ проводился по следующим сочетаниям: глагол + существительное, глагол + наречие, деепричастие + существительное, деепричастие + наречие, причастие + существительное., причастие + наречие, прилагательное + существительное, существительное + существительное. Для подробного анализа были выбраны три параметра: «существительное + существительное»,

«глагол + существительное» и «деепричастие + наречие». Целью исследования являлась проверка возможности классификации текстов по стилевым характеристикам при помощи частотных характеристик указанных выше параметров.

Первый параметр – «существительное + существительное в род.п.» (генетивная конструкция, часто соответствующая неоднословному термину). Большое количество этих конструкций традиционно рассматривается как, с одной стороны, морфолого-синтаксическая характеристика научных текстов, а с другой стороны – образчик плохо воспринимаемого стиля и поле работы редактора. По результатам экспериментов художественные коллекции содержали конструкции данного типа от 1 до 1,7% от общего числа выделенных конструкций, новостные тексты – от 18 до 29% и научные тексты – от 35 до 43% конструкций. Все коллекции оказались хорошо отделимы друг от друга по данному параметру.

В качестве второго параметра рассматривалась доля конструкций «глагол + существительное». Второй параметр интересен сам по себе и в сопоставлении с первым. Большое количество конструкций «глагол + существительное» характеризует динамические тексты, т.е. тексты, в которых реализуется большое число ситуаций. Этот параметр взаимодополняет второй, т.к. обилие генетивных конструкций, напротив, отличает статические тексты, т.е. тексты, в которых сообщается о некотором положении дел. В первом случае – рассказ о событиях, во втором – называние события. Именно поэтому для восприятия легче тексты с конструкциями «глагол + существительное».

Эксперименты показали, что в художественных текстах выделяется стабильно около 57–58% от всех конструкций данного типа. Новостная лента показывала 31–40% конструкций «глагол +существительное», научные тексты – от 20 до 28%. Таким образом, разделение текстов по жанрам также оказалось хорошим. Результаты приведены в Табл. 10.

*Таблица 10*

**Результаты экспериментов для разных параметров**

| Корпус | гл+сущ | сущ+сущ | гл+сущ/сущ+сущ |
|---|---|---|---|
| РБК | 0,25 – 0,4 | 0,35 – 0,6 | 0,4 – 1,2 |
| Лента | 0,3 – 0,45 | 0,35 – 0,5 | 0,5 – 2 |
| Пратчетт | 0,5 – 0,65 | 0,05 – 0,2 | 3 – 10 |
| Диссертация 1 | 0,2 – 0,3 | 0,4 – 0,5 | 0,4 – 0,7 |
| Диссертация 2 | 0,1 – 0,2 | 0,6 – 0,8 | 0,1 – 0,3 |
| RCDL | 0,15 – 0,3 | 0,5 – 0,7 | 0,3 – 1,2 |

В качестве третьего параметра были взяты конструкции «деепричастие + наречие». Такого типа конструкции встречаются в текстах сравнительно редко. Они, вероятно, характеризуют те тексты, в которых делается особый акцент на

действиях (оценивается «качество действия»). Можно предположить, что эти конструкции будут маркировать наиболее яркие тексты, реализуя функции воздействия на адресата.

Для художественных текстов процент подобных конструкций составил около 0,5%. Научные и новостные коллекции по данному параметру разделить не удалось. Для научных коллекций параметр принимал значения от 0,02 до 0,09%, для новостей – от 0,06 до 0,1%. Таким образом, по данному параметру отделима только часть текстов.

В результате анализа промежуточных результатов классификаций, полученных на основании трех простых параметров, был выбран один комбинированный параметр, максимально отражающий соотношение динамичности / статичности текстов коллекции. Четвертый параметр в числителе имеет частеречные сочетания, характеризующие динамичность текста, а в знаменателе – его статичность: «((гл+сущ) + (гл+нар) + (деепр+сущ) + (деепр+нар)) / ((сущ+сущ) + (прил+сущ))».

Эксперименты показали, что в художественных текстах подобное отношение принимает значения от 2,16 до 2,2; в новостных текстах от 0,67 до 0,83 (за исключением текстов смешанной тематики, где значение составило 1,38); для научных текстов были получены значения от 0,29 до 0,53. Таким образом, и здесь жанровое разделение текстов было произведено успешно.

Помимо исследования всего корпуса было проведено исследование указанных параметров в плавающем окне размером примерно 2 – 3 страницы. Как видно из рисунков 1 – 3, беллетристика хорошо отделяется от остальных корпусов даже при использовании окна сравнительно небольшого размера. При этом для новостей и научных текстов происходит пересечение области значений. В связи с этим идентификация стиля текста небольших фрагментов может быть затруднена.

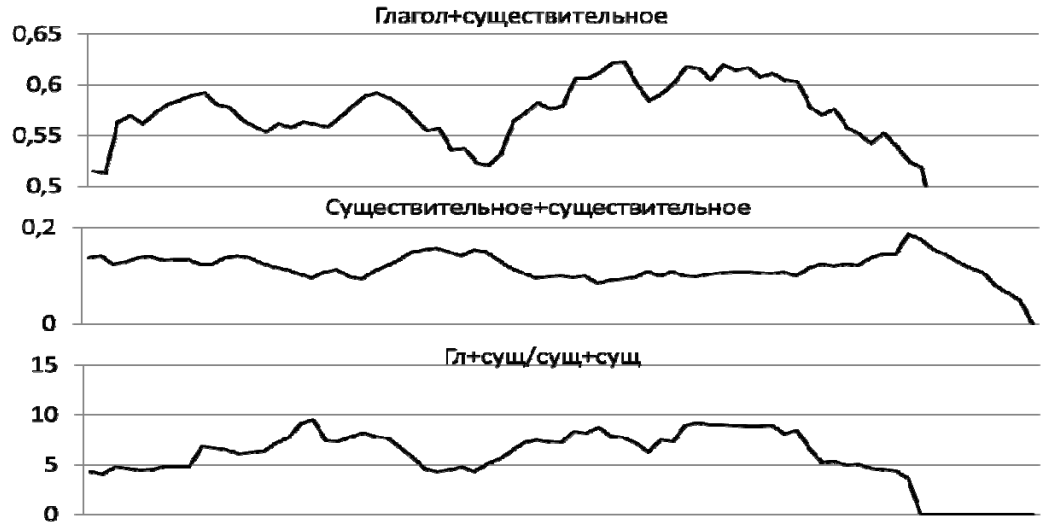


*Рис. 1.* Результаты экспериментов для Т. Пратчет «Опочтарение»

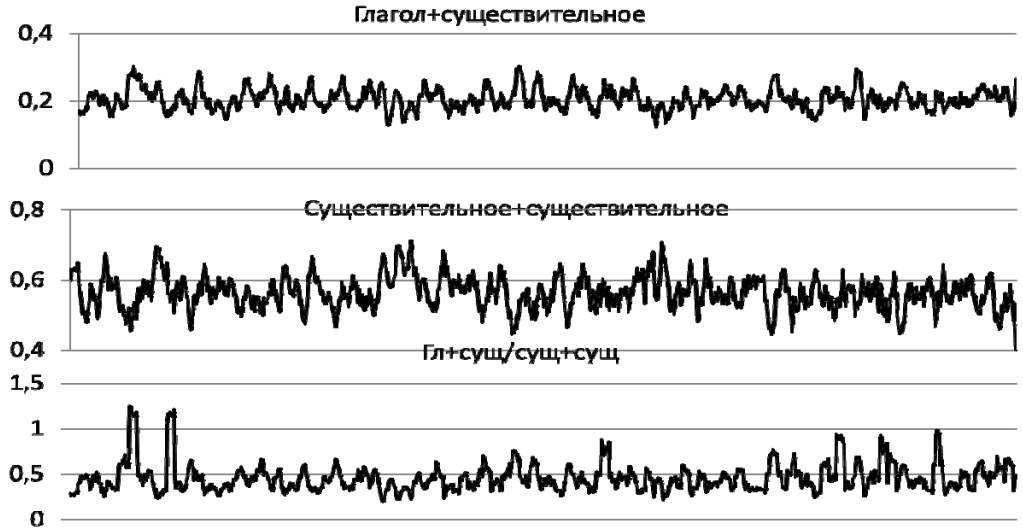


*Рис. 2.* Результаты экспериментов для материалов конференции RCDL

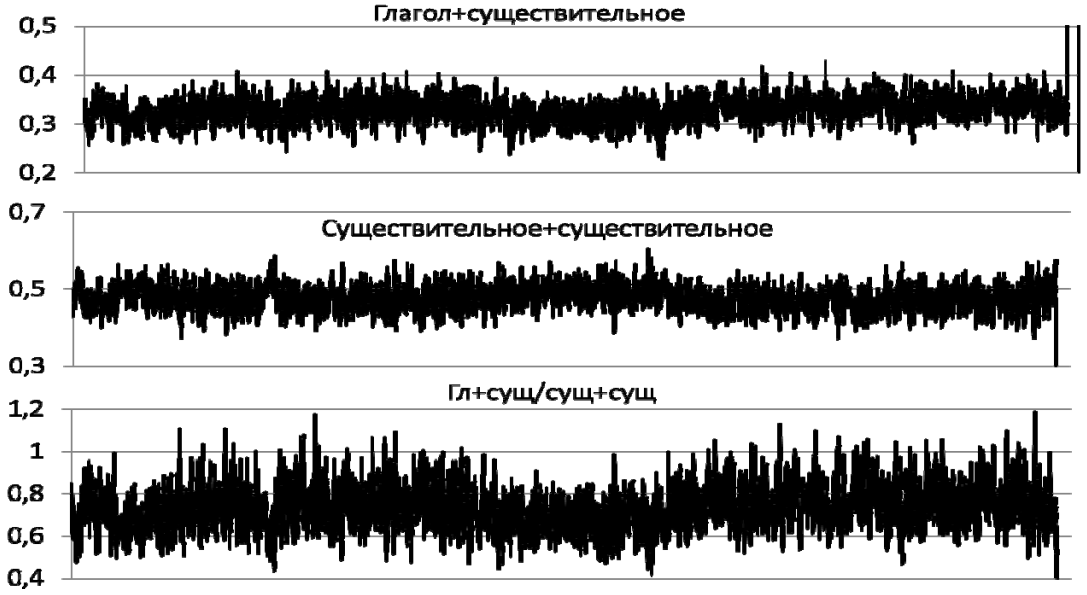


*Рис. 3.* Результаты экспериментов для материалов новостей РБК

Более того, эксперименты показали, что значения у различных источников одного стиля пересекаются, но не совпадают полностью. Это также вносит определенные трудности в определение стиля текста.

Был проведен сравнительно подробный анализ возможностей использования трех простых (некомбинированных) параметров: определение результирующего разбиения на классы; границы классов коллекций, которые ведут себя неоднозначно по отношению к используемому параметру; набор коллекций, представляющих наиболее однородную выборку с точки зрения стиля и жанра.

Было выявлено, что наиболее неоднозначные результаты показывают тексты со смешанным типом. Например, корпус текстов «It happens»,

содержащий в себе околокомпьютерные истории, чаще всего оказывался на границе между новостями и беллетристикой.

Выявленная неоднозначность может быть интерпретирована двояко:

- рассматриваемая коллекция содержит тексты, разнородные по стилевым характеристикам,
- тексты этой коллекции реализуют смешение стилей (жанровую неоднородность) внутри самого текста.

Разрешить неоднозначность интерпретации можно лишь в результате включения в наше исследование новой единицы анализа, т.е. рассмотрение распределения значений параметров по текстам в рамках указанных коллекций.

Коллекции, представляющие наиболее однородную стилевую выборку для новостных и научных текстов:

- по первому параметру – для новостных текстов это издания «Мембрана», «Компьюлента», «PCWeek», «Российская газета» и «РБК»; для научных – сборник конференции RCDL, журнал «Программные продукты и вычислительные системы», а также подборки авторефератов;
- по второму параметру – для новостных это «Мембрана», «Компьюлента», «Популярная механика»; для научных – подборки авторефератов.

Согласно рассмотренным интегральным параметрам, Мембрана и Компьюлента формируют однородную по стилю коллекцию. Является ли это характеристикой коллекций в целом? Априори можно было предположить скорее смешение жанров и стилей в текстах, одновременно характеризующихся новостной и научно-популярной природой в рамках текстов этой коллекции. И этот вопрос надо решать через анализ распределений значений параметров по текстам в рамках указанных коллекций.

Также следует заметить, что результаты могут отображать особенности использованного метода. Так, например, для глагола «думать» в научных текстах было выделено гораздо меньше конструкций. Это связано с тем, что в художественных текстах чаще всего употребляются фразы вида «думать что-то о чем-то», тогда как в научных текстах более распространены конструкции вида «мы думаем, что», которые не могли быть приняты в рассмотрение.

Несмотря на это, метод показал свою пригодность для классификации коллекций текстов по различным жанрам. В дальнейшем мы планируем изучить статистические характеристики отдельных текстов или их фрагментов, распределение частот встречаемости различных конструкций по элементам текста равного размера.

## Метод генерации базы данных глагольного управления

*Шаг 3* – составление модели управления предлогов. Из большой базы сочетаний, полученных на Шаге 1, отбираются те, в которых существительное однозначно по падежу. Для них рассчитывается статистика встречаемости пар вида «предлог + существительное в заданном падеже». Для удобства будем считать, что все пары с одним предлогом собираются в единую запись. Всего была получена информация о примерно 300 предлогах (в том числе и составных). Таким образом, для нашего умозрительного корпуса получим записи следующего вида:

```
К 0*0*8950*21*17*5
НА 0*0*0*30707*0*89
ПЕРЕД;0*0*0*0*5*0
```

Цифры показывают, сколько раз данный предлог встретился с существительными в именительном, родительном, дательном, винительном, творительном и предложном падежах соответственно. Цифры здесь взяты от глагола «приглашать», но в ходе экспериментов бралась информация по всем сочетаниям, в которых использовался данный предлог.

Из модели управления предлогов вручную был отсеян шум, связанный с ошибками в тексте или методе выделения. Анализировалась частотная информация, полученная для предлогов, а не вся модель управления в целом. Так, в нашем примере предлог «К» никогда не сочетается с существительными в винительном, творительном или предложном падежах. Таким образом, мы получили информацию о том, в каком падеже может встречаться существительное, связанное с данным предлогом.

```
К 0*0*8950*0*0*0
НА 0*0*0*30707*0*89
ПЕРЕД;0*0*0*0*5*0
```

Заметим, что также здесь отсеивались сочетания предлога с именительным падежом. Как показала практика, несмотря на то, что подобное сочетание возможно (например, «идти в солдаты» или предлог «а-ля», требующий после себя именительного падежа), но в получаемой базе в подавляющем большинстве случаев является шумом. Таким образом, мы жертвуем полнотой результатов в пользу их корректности.

*Шаг 4* – получение модели управления для глаголов. Из базы, полученной на шаге 1, строим базу сочетаний вида «глагол+предлог+падеж существительного», после чего отбрасываем все варианты, запрещенные моделью предлога, полученной на шаге 3, или встречающиеся только один раз (шаг 2). Вновь для удобства объединим информацию о паре «глагол+предлог» в единую запись в том же формате. Подобным образом будут получены записи следующего вида.

```
ПРИГЛАШАТЬ;17063*2103*863*14439*1583*41
ПРИГЛАШАТЬ;БЕЗ;0*97*0*0*0*0
ПРИГЛАШАТЬ;В;0*89*0*17847*0*1775
ПРИГЛАШАТЬ;ВМЕСТО;0*9*0*0*0*0
ПРИГЛАШАТЬ;ВО;0*0*0*1005*0*29
ПРИГЛАШАТЬ;ВОЗЛЕ;0*5*0*0*0*0
ПРИГЛАШАТЬ;ВОПРЕКИ;0*0*9*0*0*0
ПРИГЛАШАТЬ;ДЛЯ;0*1129*0*0*0*0
ПРИГЛАШАТЬ;ДО;0*1*0*0*0*0
ПРИГЛАШАТЬ;ЗА;0*0*0*817*25*0
ПРИГЛАШАТЬ;ИЗ;0*297*0*0*0*0
ПРИГЛАШАТЬ;ИЗ-ЗА;0*61*0*0*0*0
ПРИГЛАШАТЬ;ИЗО;0*5*0*0*0*0
ПРИГЛАШАТЬ;К;0*0*8950*21*17*5
ПРИГЛАШАТЬ;КО;0*0*489*0*0*0
ПРИГЛАШАТЬ;КРОМЕ;0*9*0*0*0*0
ПРИГЛАШАТЬ;НА;0*0*0*30707*0*89
ПРИГЛАШАТЬ;НАД;0*0*0*0*21*0
ПРИГЛАШАТЬ;ОТ;0*17*0*0*0*0
ПРИГЛАШАТЬ;ПЕРЕД;0*0*0*0*5*0
ПРИГЛАШАТЬ;ПО;0*0*509*0*0*65
ПРИГЛАШАТЬ;ПОД;0*0*0*117*19*0
ПРИГЛАШАТЬ;ПОДОБНО;0*0*5*0*0*0
ПРИГЛАШАТЬ;ПОСЛЕ;0*189*0*0*0*0
ПРИГЛАШАТЬ;ПОСРЕДИ;0*9*0*0*0*0
ПРИГЛАШАТЬ;ПРИ;0*0*0*0*0*1
ПРИГЛАШАТЬ;РАДИ;0*17*0*0*0*0
ПРИГЛАШАТЬ;С;0*0*1*0*815*0
ПРИГЛАШАТЬ;СО;0*1*0*0*18*0
ПРИГЛАШАТЬ;СОГЛАСНО;0*0*13*0*0*0
ПРИГЛАШАТЬ;ЧЕРЕЗ;0*0*0*109*0*0
```

Эта информация и является искомой базой данных моделей глагольного управления. Здесь приведена вся информация, полученная для глагола «приглашать». Фильтрация результатов для присоединения существительного без предлога не проводилась. Это связано с тем, что для автоматической фильтрации отношения цифр оказалось недостаточно. Для некоторых глаголов стандартные ошибки оказывались более частотными, чем корректные варианты. Анализ полного списка глаголов вручную при этом не проводился, хотя и возможен (обработка 25000 – 30000 глаголов может быть проведена за приемлемое время).

В приведенных выше сочетаниях есть несколько вхождений, которые выглядят неочевидно, и их необходимо проиллюстрировать примерами. Так, скажем, «приглашать кроме» выглядит корректно только в сочетаниях вида

«Кроме капитана приглашали других гостей», хотя в самом примере отсутствует запятая, то есть выделение происходит с ошибкой. Предлог «при» в данном случае обозначает условие приглашения, например, «приглашать при болезни врача». А предлог «над» встретился в сочетании «приглашать над жизнью», что, скорее всего, является ошибкой в тексте.

*Шаг 5* – фильтрация базы сочетаемости. Из базы, полученной на шаге 1, отсеиваем все сочетания, не подходящие под полученную на шаге 4 модель управления. Одновременно может быть устранена падежная неоднозначность. В итоге мы получаем базу сочетаемости глаголов с существительными.

## Результаты экспериментов

Нами был взят корпус объемом 10,5 млрд. словоупотреблений. Корпус состоял из нескольких подкорпусов беллетристики (Библиотека Мошкова – 688 млн. словоупотреблений, lib.rus.ec – 8,9 млрд.), новостных текстов за 1999 – 2011 гг. (самые значимые РИА Новости – 220 млн., Коммерсант и Независимая газета – по 99 млн., Взгляд – 95 млн., общий объем – около 800 млн.), научных текстов (авторефераты, диссертации, статьи из сборников конференций и журналов, объемом несколько десятков миллионов словоупотреблений, общий объем – более 60 млн. словоупотреблений).

Из данного корпуса было выделено более 23 млн. уникальных связок вида «глагол+предлог+существительное» без учета формы существительного и глагола. Из полученной базы применением трех шагов описанного алгоритма было извлечено около 425 000 сочетаний «глагол + предлог + разрешенные падежи». Качество полученных результатов находится на уровне не ниже 95% для сочетаний «глагол + предлог + разрешенные падежи» и около 99% для «глагол + предлог + падеж» для однословных предлогов. Качество определения составных предлогов несколько ниже, так как слова, входящие в составной предлог, могут использоваться и в качестве именной группы. Оценка пропусков в модели не проводилась. Для разметки корпуса использовалась система морфологического анализа «Кросслейтор» [Елкин 2003, Школа 2011].

Заметим, что в полученный словарь глагольного управления не попала значительная часть слов. Исследование результатов показало, что в выделенных связках «глагол+предлог+существительное» приняло участие 22200 глаголов из 26400, представленных в использованном морфологическом словаре, и 55200 из 83000 существительных. Слова не были включены в словарь по одной из двух причин. Во-первых, это редко встречающиеся слова, имеющиеся в морфологическом словаре, например, «парагвайка», «восьмиугольный» и т.д. Во-вторых, это существительные, омонимичные прилагательным во всех своих формах, например, «белый», «красный», «больной». Заметим, что введение шаблонов 5 и 6 позволило сократить количество подобных слов.

Кроме того, полученные сочетания совершенно не различают лексических значений глаголов. Несмотря на это, наличие подобного словаря сочетаний

позволяет перейти к практическому решению целого ряда задач в анализе текстов: снятия омонимии, фильтрации результатов синтаксического анализа и др. Кроме того, наличие большой базы позволит попытаться перейти и к задачам, связанным с семантикой.

Сравнение новых результатов с предыдущими [Клышинский 2011], полученными на корпусе в 7,5 млрд словоупотреблений, показывает, что происходит постепенное насыщение словаря. Так, при увеличении корпуса примерно на 30% число глагольных сочетаний выросло лишь примерно на 15%. Аналогичная ситуация наблюдается и для других сочетаний. Кроме того, были проведены эксперименты на корпусе большего объема, содержащем лишь новостные тексты. Если сравнивать результаты, полученные на этом корпусе, с результатами, полученными на художественных текстах, полученных с использованием lib.rus.ec, то можно увидеть, что при вдвое большем объеме новостного корпуса на нем было получено 4,1 млн сочетаний «существительное + прилагательное» против 5,4 млн, полученных на художественных текстах. Аналогичная ситуация наблюдается и с комбинациями «глагол + существительное»: 14,3 млн комбинаций против 21,6 млн. Таким образом, можно сделать вывод, что для извлечения МУ из текстов более важны стиль изложения и богатство лексики, чем объем корпуса.

Для оценки было выбрано по 1 – 2 тысячи конструкций, выделяемых из текста, которые были просмотрены информантами-носителями языка (часть из них имела специальное образование). Каждый фрагмент просматривался двумя информантами. В итоге было выяснено, что количество ошибок в выделяемых фрагментов примерно следующее – существительное + прилагательное – менее 1%, глагол + предлог + существительное – 4 – 7%, существительное + существительное в родительном падеже – до 5%. При этом несколько ошибок, принадлежащих одному виду, повторяется многократно, и их доля может быть уменьшена за счет введения дополнительных фильтров в предложенный метод. Кроме того, так как оценивался материал, извлекаемый непосредственно из текста, результаты в базе сочетаемости будут несколько лучше, так как часть сочетаний повторяется. Таким образом, сочетания существительного и прилагательного могут использоваться для практических приложений уже сейчас, а сочетания глагола и существительного и существительного и существительного в родительном падеже требуют дальнейшей доработки.

Для сравнения результатов был использован словарь [Бирюк 2012]. Из полученных сочетаний и указанного словаря были выбраны глаголы, сочетающиеся со словом «пример» без предлога. В связи с большим количеством сочетаний, полученных в базе, анализ проводился только для глаголов на буквы А, Б и В. В словаре [Бирюк 2012] было найдено только два сочетания: «брать пример» и «взять пример». В полученном по предложенному методу словаре было найдено 139 сочетаний, из них 104 встретились более одного раза, а 77 – больше двух. Десять самых частотных сочетаний связаны со следующими глаголами: быть (встретилось 18706 раз), видеть (1204 раз),

воспользоваться (1043 раза), брать (785 раз), вспомнить (726 раз), бывать (532 раза), взять (468 раз), вдохновлять (418 раз), встречаться (403 раза), вдохновляться (192 раза). Данный пример показывает, что полнота полученного словаря сочетаний значительно выше. Из проанализированных сочетаний сомнения вызвало только одно «ВЕЛЕТЬ ПРИМЕР» (по всей видимости «велеть примером показать что-то»). Однако помимо этого было обнаружено некоторое количество сочетаний глагола с «на пример» (искаженное «например»), количество которых не превысило 1% от всех сочетаний со словом «пример». Заметим, что подобное сочетание является синтаксически корректным, хотя и грамматически неверным.

В заключение заметим, что описанный метод был применен для текстов на английском языке [Гурбанов 2012] и показал хорошие результаты. При применении методов снятия омонимии он показывает сопоставимую полноту выделения сочетаний, хотя это несколько понижает качество результатов.

## Список литературы

**[Апресян 1982]** Апресян Ю.Д., Палл Э. Русский глагол – венгерский глагол. Управление и сочетаемость. Будапешт, 1982.

**[Бирюк 2012]** Бирюк О.Л., Гусев В.Ю., Калинина Е.Ю. Словарь глагольной сочетаемости непредметных имен русского языка // электронная публикация, URL: http://dict.ruslang.ru/abstr_noun.php

**[Большаков 2011]** Большаков И.А., Гельбух А.Ф. Большой электронный словарь как политематический справочник и формирователь запросов к Интернету // Материалы международной конференции «Диалог 2011», 2011 г. С. 124–134.

**[Волкова 2003]** Волкова И.А., Мальковский М.Г., Одинцев Н.В. Адаптивный синтаксический анализатор // Диалог 2003: Труды Международного семинара. М., 2003. Т. 1. С. 401–406.

**[Гельбух 1999]** Гельбух А. Разрешение синтаксической неоднозначности и извлечение словаря моделей управления из корпуса текстов // Материалы VIII Международной конференции KDS-99 (Крым - 13-18.09.1999г. - Кацивели)

**[Герд 1996]** Герд А.С. Специальный текст как предмет прикладной лингвистики / А.С. Герд // Прикладное языкознание : учебник / отв. ред. А.С. Герд. – СПб. : Изд-во С.-Петерб. ун-та, 1996. С. 68–90.

**[Гурбанов 2012]** Гурбанов Т.П., Клышинский Э.С. Параллельный алгоритм составления словаря глагольного управления для новостных текстов на английском языке // Сб. трудов 15 научно-практического семинара «Новые информационные технологии», М., 2012. С. 203–212.

**[Денисова 2002]** Словарь сочетаемости слов русского языка / Под ред. П.Н. Денисова, В.В. Морковкина. 3-е изд., испр. М., АСТ, 2002. 816 с.

**[Елкин 2003]** Елкин С.В., Клышинский Э.С., Стеклянников С.Е. Проблемы создания универсального морфосемантического словаря // Сб. трудов

Международных конференций IEEE AIS'03 и CAD-2003, т. 1, Дивноморское. 2003

**[Клышинский 2011]** Клышинский Э.С., Кочеткова Н.А., Литвинов М.И., Максимов В.Ю. Метод разрешения частеречной омонимии на основе применения корпуса синтаксической сочетаемости слов в русском языке // Научно-техническая информация. Сер. 2: Информационные системы и процессы. №1 2011 г., С. 31–35

**[Кобрицов 2007]** Кобрицов Б.П., Ляшевская О.Н., Толдова С.Ю. Снятие семантической многозначности глаголов с использованием моделей управления, извлеченных из электронных толковых словарей // Интернет-математика – 2007, URL: http://download.yandex.ru/IMAT2007/kobricov.pdf

**[Ляшевская 2009]** Ляшевская О.Н., Кузнецова Ю.Л. Русский фреймнет: к задаче создания корпусного словаря конструкций // Труды конфренции «Диалог 2009». М.: 2009. С. 306–312.

**[НКРЯ 2012]** Статистика. Национальный корпус русского языка // Электронная публикация, режим доступа – http://ruscorpora.ru/corpora-stat.html

**[Розенталь 1986]** Розенталь Д.Э. Управление в русском языке // М.: Книга, 1986 г.

**[СинТагРус 2011]** Итоговый отчет о работе по программе Фундаментальных исследований Президиума РАН «Корпусная лингвистика» 2011 г. «Разработка системы синтаксического анализа русских текстов на базе корпуса «СинТагРус»» URL: http://corpling-ran.ru/files/I/boguslavsky.pdf

**[Толдова 2008]** Толдова С.Ю., Кустова Г.И., Ляшевская О.Н. Семантические фильтры для разрешения многозначности в национальном корпусе русского языка: глаголы // Труды конфренции «Диалог 2008». М., 2008. С. 522–529.

**[Школа 2011]** Материалы Летней студенческой школы компьютерной лингвистики. URL: http://clschool.miem.edu.ru/Материалы-школы.html

**[Framebank 2012] С**истема Fraemebank. URL: http://framebank.ru/

**[Frolova 2011]** Frolova T.I., Podlesskaya O.Yu. Tagging lexical functions in Russian texts of SynTagRus // Материалы международной конференции «Диалог 2011», 2011 г. С. 207–218

**[Manning 1993]** Automatic Acquisition of a Large Subcategorization Dictionary from Corpora // In Proc. of the 31st Meeting of ACL, pp. 235–242.

**[Messiant 2008]** Messiant C., Korhonen F., Poibeau T. LexSchem: A Large Subcategorization Lexicon for French Verbs // In Proc. of "LREC 2008, Marrakech: Morocco (2008)"

**[Preiss 2007]** Preiss J., Briscoe T., Korhonen A. A System for Large-Scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from Corpora // in Proc. of the 45 Annual Meeting of the Association of Computational Linguistics, pp. 912-919.

**[Pymorphy 2012]** Pymorphy v0.5.5 documentation. URL: http://packages.python.org/pymorphy/usage/base.html

# Оглавление