Анализ омонимичных словосочетаний, эквивалентных слову

К.К. Боярский[1], Е.А. Каневский [2], Е.Н. Клименко [2], Е.Ф. Силина [2]

[1] Университет ИТМО

[2] Санкт-Петербургский экономико-математический институт РАН

boyarin9@yandex.ru, kanev@emi.nw.ru

Аннотация

Работа посвящена анализу словосочетаний, эквивалентных слову. Рассматриваются способы снятия омонимии у словосочетаний, с одной стороны являющихся устойчивыми оборотами, выполняющими функции наречий и предлогов, а с другой — представляющих простое сочетание предлога с существительным. Утверждается, что анализа ближайшего контекста часто недостаточно для решения вопроса о том, является ли словосочетание устойчивым оборотом.

В зависимости от глубины использования контекста можно выделить 4 группы словосочетаний. Для снятия омонимии у словосочетаний первой группы достаточно проанализировать морфологические признаки слов ближайшего контекста, второй группы — проверить наличие определенных лемм или словоформ, третьей — проанализировать семантические классы окружающих слов, четвертой — проверить комплекс признаков у слов удаленного контекста. Отдельно рассматриваются особые словосочетания, имеющие три варианта разбора: два оборота, соответствующие разным частям речи и простое сочетание предлога с существительным.

Показано, что при автоматическом разборе русскоязычного текста можно подобрать правила, позволяющие достаточно надежно снимать омонимию у подобных словосочетаний. Приводятся примеры, показывающие значение удаленного контекста для анализа оборотов.

Ключевые слова: анализ предложения, словосочетания, лексические обороты, морфологический анализ, части речи, леммы, семантические классы.

1. Введение

При автоматическом разборе предложений русского языка и построении дерева зависимостей возникает проблема снятия морфологической неоднозначности. Одним из путей ее решения является широкое использование стандартных сочетаний слов. Подробная классификация таких сочетаний приведена в [1]. Нас будут, прежде всего, интересовать такие словосочетания, которые занимают промежуточное положение между простыми коллокациями, сохраняющими семантику входящих в них слов, и фразеологизмами (идиомами), в которых исходное значение полностью меняется. Большая часть неизменяемых словосочетаний являются оборотами, выполняющими функции:

- наречий — *в конце концов, время от времени, в свое время, на самом деле*;
- предлогов — *в зависимости от, в соответствии с, несмотря на, по мнению*;
- вводных оборотов — *другими словами, к слову сказать, по вашему мнению*;
- союзов — *а также, в том числе и, вместо того чтобы, если бы, как тольк*о;
- частиц — *все же, как бы, как раз, к тому же, едва не, вроде бы, вроде как*;
- предикативных оборотов — *не дай бог, лыка не вяжет, нечего было делать*.

Наиболее полные списки таких оборотов приведены в НКРЯ [2]. Для дополнительной информации мы использовали словари С.А. Кузнецова [3] и Р.П. Рогожниковой [4].

Как показывает анализ существующих синтаксических и семантических парсеров, сегодня существует два подхода к разбору подобных оборотов. Первый подход не предполагает какого-либо специального графематического их выделения — парсер «Этап-3» [5]. При втором подходе такой оборот выделяется особым образом — парсер фирмы ABBYY [6].

В нашей работе для синтаксического разбора предложения использовался парсер SemSin [7] [8], который такой оборот рассматривает как один узел (токен) и не разбивает его на отдельные словоформы, как это видно на примере разбора предложения *Другими словами, продажа производится в зависимости от качества продукции* (см. рис. 1). Здесь словосочетание *другими словами* является вводным оборотом, а словосочетание *в зависимости от* выполняет функции предлога.

Более сложная ситуация возникает тогда, когда сочетание нескольких слов в зависимости от контекста может быть оборотом, а может и не быть им. Многие словосочетания такого рода рассмотрены Р.П. Рогожниковой [4], которая отмечает возможность их использования в качестве свободных словосочетаний, омонимичных оборотам. Однако нас интересует не только сам факт их наличия, но и возможность их максимально правильного опознания при автоматическом анализе предложений. Ситуация осложняется тем, что это решение необходимо проводить на этапе токенизации после морфологического анализа, но до начала синтаксического разбора.

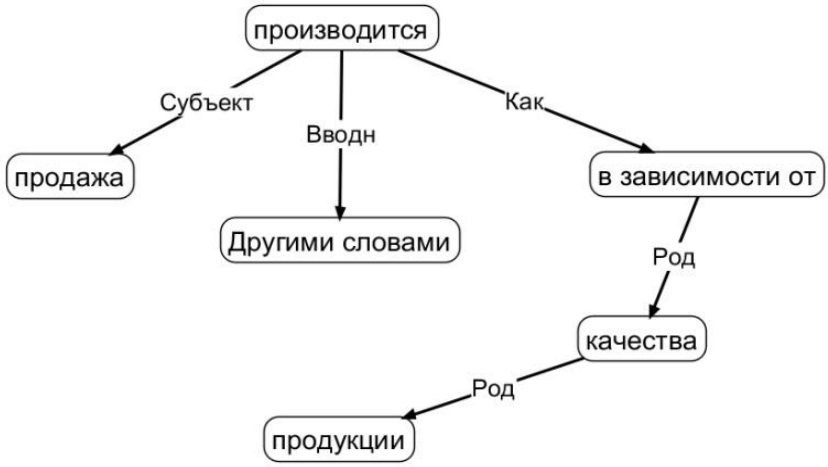

Рис. 1. Дерево разбора

Мы рассматривали каждое такое словосочетание отдельно, анализируя по 100–200 отобранных из НКРЯ предложений. Использовались также предложения, выделенные из дополнительного корпуса, составленного из нескольких повестей и новостных текстов, объемом более 60 млн. слов. Отдельные словосочетания такого рода уже были рассмотрены нами ранее [9] [10] [11]. Ниже обсуждаются особенности алгоритмов, позволяющих решить вопрос: является ли данное словосочетание оборотом или нет. Для решения этого вопроса необходимо проанализировать окрестности исследуемого словосочетания. Для большей части словосочетаний достаточно анализа **ближайшего контекста**, который мы определим в 2 слова слева и 3 слова справа от собственно словосочетания. Фактически производится поиск определенных существительных или предикатов, которые могут быть отделены от рассматриваемого словосочетания прилагательными или наречиями. Следует заметить, что анализ словосочетаний производится на самом раннем этапе разбора текста, когда именные группы еще не образованы. Для некоторых словосочетаний необходим анализ **удаленного контекста**, который включает в себя пространство до 10 слов влево и до 12 слов вправо от собственно словосочетания.

Рассмотрим подробнее отобранные нами обороты, выполняющие функции наречий и предлогов.

2. Наречные обороты

В используемом нами словаре содержится около 880 наречных оборотов (в НКРЯ — около 2220 наречных и предикативных оборотов). В том числе, около 130 наречных оборотов являются омонимичными [4]. Мы определили правила разрешения омонимии для 57 таких словосочетаний, которые сведены в табл. 1 в пять групп в зависимости от используемой информации.

По умолчанию считается, что в ближайшем контексте анализируются те или иные морфологические признаки. Дополнительно в ближайшем контексте может производиться поиск определенных словоформ или лемм (пометка в столбце «Б С»). По сути дела, для каждого словосочетания используются своего рода микрословари. В ближайшем контексте может также осуществляться поиск определенных семантических классов (пометка в столбце «Б Кл»). Классы слов определяются в соответствии с описанным впервые в [12] и впоследствии модифицированным нами классификатором на 1700 классов, входящим в состав словарного обеспечения парсера SemSin. Наконец, может проводиться поиск словоформ, лемм, предикатных элементов или слов с заданным классом в удаленном контексте (пометка в столбце «УД С»).

Таблица 1. Омонимичные наречные обороты по группам

	Словосочетания	**Б С**	**Б Кл**	**УД С**
1	В БОЛЬШИНСТВЕ, В ЖИЗНЬ СВОЮ, В ОТКРЫТУЮ, ВДОЛЬ И ПОПЕРЕК, ДЛЯ НАЧАЛА, ДО КОНЦА, ПО ВИДУ, ПО ИМЕНИ, ПО НАЗВАНИЮ, ПО НАРАСТАЮЩЕЙ, ПО ОТЧЕСТВУ, ПОД КОНЕЦ, ПОСЛЕ ЭТОГО			
2	В ГЛАВНОМ, В ДАЛЬНЕЙШЕМ, В МОМЕНТ, В ОСНОВНОМ, В ПРОШЛОМ, В РЕЗУЛЬТАТЕ, В РОСТ, В СОВЕРШЕНСТВЕ, В СРЕДНЕМ, В СРОК, В ЦЕЛОМ, ВСЕ ЕЩЕ, ВСЕМИ СИЛАМИ, КАК ЕСТЬ, КАК ПОПАЛО, КАК СЛЕДУЕТ, КАК СОБАКА, ПЕРЕД ТЕМ	+		
3	БЕЗ КОНЦА, БЕЗ ТРУДА, В ИТОГЕ, ДО СМЕРТИ, ЗА ГРАНИЦЕЙ, ЗА ГРАНИЦУ, НА ПАРУ, НА ЭТОТ СЧЕТ, ПО ФАМИЛИИ, С КОРНЕМ, СО ВСЕХ КОНЦОВ	+	+	
4	В БУДУЩЕМ, В ЗАКЛЮЧЕНИЕ, В ЛОБ, В НОЧЬ, В ОСОБЕННОСТИ, В ОТВЕТ, ДЛЯ ФОРМЫ, ДО БОЛИ, ДО СЛЕЗ, ЗА РУБЕЖ, НА СВЕТ, ОДИН РАЗ, ОКОЛО ТОГО, С ГОДАМИ	+	+	+
5	БЕЗ ВЕСТИ, ЗА РУБЕЖОМ, КАК ПРЕЖДЕ			

В первую группу входят наречные обороты, для анализа которых используются только морфологические признаки словоформ в ближайшем контексте. Типичным их представителем является словосочетание ПОД КОНЕЦ.

Как правило, это наречие[1]: ● *И выясняется* ***под конец****, что и впрямь делить-то нечего.*

Однако это словосочетание может быть сочетанием предлога и существительного, если за ним следует слово в родительном падеже: ● *Под конец* <u>*моего*</u> *рассказа он стал кивать головой, не отнимая рук от лица*.

[1] Здесь и далее примеры отмечены знаком “●”, жирным шрифтом выделены словосочетания, являющиеся оборотами, подчеркнуты слова, позволяющие отличить словосочетания, не являющиеся оборотами.

Ко второй группе относятся наречные обороты, для анализа которых дополнительно требуется проверить наличие определенных словоформ или лексем в ближайшем контексте. Типичным их представителем является словосочетание В ЦЕЛОМ.

Как правило, это наречие: ● *Нельзя сказать, что ИКТ **в целом** до сих пор не интересовали инвесторов.*

Может быть сочетанием предлога и прилагательного, если за ним следуют существительное или прилагательное в предложном падеже: ● *Она думала, что одна в целом мире*.

Это существительное может следовать за оборотом не контактно, а отделяться словами *его, ее, их*: ● *Отсылаю заинтересованного читателя к тексту данной статьи в целом ее виде…*

К третьей группе принадлежат те наречные обороты, для анализа которых в ближайшем контексте кроме морфологических признаков дополнительно анализируются семантические классы. Примером может служить словосочетание С КОРНЕМ.

Это наречие, если непосредственно перед ним находится одна из словоформ от лемм *вывернуть, выворотить, выдернуть, вырвать:* ● *Ветром выворачивало **с корнем** столетние дубы.*

Однако может быть сочетанием предлога и существительного, если за ним следует существительное в родительном падеже единственного числа из класса «Растения» (например, *алтей, валериана, девясил*): ● *Я пил чай с корнем солодки;*

или слово в кавычках: ● *Ещё интересны на Балканах топонимы с корнем "козак"*;

или слово, оканчивающееся на дефис или начинающееся с заглавной буквы, или написанное латинскими буквами: ● *Сытно перекусив, духи скажут, что сделать с корнем Ронго.*

В противном случае это наречие: ● *Если этих людей выловить и расстрелять после войны, военная мощь Германии будет уничтожена с корнем*.

В четвертую группу включены наречные обороты, для анализа которых необходимо использовать удаленный контекст. Рассмотрим словосочетание ДЛЯ ФОРМЫ.

Это наречие, если непосредственно перед словосочетанием находится одно из слов *больше, даже, лишь, разве, словах, только, хотя, это:* ● *Гарнизоны стояли кое-где только **для формы***;

или непосредственно перед или после словосочетанием находится запятая или кавычки: ● *Впрочем, к вечеру она, **для формы**, созвала опытнейших городских будочников и открыла совещание;*

или в удаленном контексте слева или справа от словосочетания находится словоформа от одной из лемм *быть, глянуть, думать, записать, писать, подписать, пригласить, сидеть, спросить, стать*: ● Запишите нам коллективное замечание **для формы**;

или в удаленном контексте слева или справа от словосочетания находится глагол из класса «Знания Сообщение Информация Речь» (например, *сказать, советоваться, переговорить*): ● *Опросили и его **для формы***.

В противном случае это сочетание предлога и существительного: ● *Материалом для формы служила липовая или грушевая доска продольного распила*.

В пятую группу отнесены наречные обороты, омонимичные в соответствии с [4], и, в принципе, могущие рассматриваться как два отдельных слова, однако последний вариант на практике встречается крайне редко.

БЕЗ ВЕСТИ — это наречие: ● *Отец был военным и пропал на войне* ***без вести.***

Теоретически может быть и сочетанием предлога и существительного: (если за ним следуют предлоги *о* или *об*): ● *Судя по дневнику, Васмус в течение года оставлял командование без вести о себе.*

Частота этого явления 0,2% (при выборке в 1900 предложений), так что практически это словосочетание можно считать однозначным наречием.

ЗА РУБЕЖОМ — это наречие: ● *Фирма периодически арестовывает то или иное российское государственное имущество* ***за рубежом.***

Возможно сочетание предлога и существительного: ● *Справочные материалы могут быть использованы для изучения военной деятельности за рубежом второй половины XX века*

Частота этого явления 0.8% (при выборке в 2200 предложений), так что практически это словосочетание можно считать однозначным наречием.

КАК ПРЕЖДЕ — это наречие: ● *От Кямала пахло не земляникой,* ***как прежде****, а тем, что он съел.*

Теоретически может быть сочетанием союза с предлогом: ● *Как прежде образования государственного тела был период, когда все бродили, так же должно быть и перед духовным рождением России.*

Частота этого явления 0.8% (при выборке в 410 предложений), так что практически и это словосочетание можно считать однозначным наречием.

3. Обороты в функции предлогов

В используемом нами словаре содержится более 190 оборотов в функции предлогов (в НКРЯ — около 315 таких оборотов). Из них около 50 оборотов являются омонимичными [4]. Мы определили правила разрешения омонимии более чем для 45 словосочетаний, которые сведены в табл. 2. Эта таблица построена по тем же принципам, что и табл. 1

К первой группе относятся обороты, для анализа которых используются только морфологические признаки словоформ в ближайшем контексте. Типичным их представителем является словосочетание В ОКРУЖЕНИИ (!Род)[2].

Это предлог, если за ним следует слово в родительном падеже: ● *Он умер в спокойной обстановке* ***в окружении*** *близких.*

Иначе это сочетание предлога и существительного: ● *В окружении оказались более миллиона бойцов, в том числе почти вся 9-ая армия*.

[2] Здесь и далее в круглых скобках указан падеж, который требует оборот в функции предлога.

Таблица 2. Омонимичные обороты в функции предлогов по группам

	Словосочетания	Б С	Б Кл	Уд С
1	БЕЗ ПОМОЩИ, В ОКРУЖЕНИИ, В РЯДУ, В ПРЕДЕЛАХ, В СЛУЧАЕ, В СФЕРЕ, В СЧЕТ, В ЧАСТИ, В ЧИСЛЕ, В ЧИСЛО, ЗА ГРАНИЦАМИ, ЗА ГРАНИЦЫ, ПО ВОПРОСАМ, ПО ВОПРОСУ, ПО ЧАСТИ, ПОД ЗНАКОМ, С ПОМОЩЬЮ, С ТОЧКИ ЗРЕНИЯ, С ЦЕЛЬЮ			
2	В РЕЗУЛЬТАТЕ, В СВЯЗИ С, ВМЕСТЕ С, НА ПОЧВЕ, СМОТРЯ ПО	+		
3	В ДЕЛЕ, В КАЧЕСТВЕ, В МАНЕРЕ, В ОТНОШЕНИИ, В РОЛИ, В СВЕТЕ, В СИЛУ, В СТИЛЕ, В УСЛОВИЯХ, ИЗ ОБЛАСТИ, НА БАЗЕ	+	+	
4	В ЗАКЛЮЧЕНИЕ, В ЛИЦЕ, В ОБЛАСТИ, В ОБЛАСТЬ, В ПАМЯТЬ, В ПОРЯДКЕ, В ПРОЦЕССЕ, В ФОРМЕ, ПЕРЕД ЛИЦОМ	+	+	+
5	В ЦЕЛЯХ			

Ко второй группе относятся обороты в функции предлога, для анализа которых дополнительно требуется проверить наличие определенных словоформ или лексем в ближайшем контексте. Типичным их представителем является словосочетание В СВЯЗИ С (!Тв).

Это предлог, если за ним следует слово в творительном падеже: ● Верховная рада Украины ужесточила уголовное наказание за сепаратизм **в связи с** событиями на юго-востоке.

Может быть сочетанием предлога, существительного и предлога, если перед словосочетанием стоит глагол *состоять* (возможно, отделенный наречием), или после него существительное одушевленное: ● *В ней значилось, что Амосова состояла в связи с врагом народа Кадацкой*.

К третьей группе принадлежат те обороты, для анализа которых в ближайшем контексте кроме морфологических признаков дополнительно анализируются семантические классы. Примером может служить словосочетание В ОТНОШЕНИИ (!Род).

Это предлог, если за ним следует слово в родительном падеже: ● *У врача не было сомнений **в отношении** успешного окончания операции.*

Однако возможно сочетание предлога и существительного, если правее него (между ними могут быть одно или два согласованных прилагательных или причастий) стоит существительное в родительном падеже из класса «Человек» (*брат, отец, сын*), за которым следует предлог *к* или *ко*. При этом между анализируемыми словами могут быть согласованные прилагательные и причастия: ● *То был трудный миг в отношении моей матери ко мне*.

В четвертую группу включены обороты в функции предлога, для анализа которых необходимо использовать удаленный контекст. Рассмотрим словосочетание В ПАМЯТЬ (!Род, !Дат).

Это предлог, если за ним следует слово в родительном или дательном падеже: ● *С тех пор римляне **в память** этого дня завели у себя праздник*.

Однако это сочетание предлога и существительного, если слово в родительном или дательном падеже принадлежит к классу «Вещь», и при этом не является названием (компьютер, телефон, сервер): ● *Информация всех каналов параллельно записывается в память компьютера*;

или в удаленном контексте слева или справа от словосочетания находится словоформа от одной из лемм *внести, вносить, войти, вплести, врезаться, входить, залезть, запасть, записываться, отправлять, попадать, попасть, ронять:* ● *Не одними жалобами на низкие зарплаты входили в историю, в память потомков благороднейшие российские учителя*;

или правее этого словосочетания отсутствует прилагательное или существительное в родительном или дательном падеже: ● *Одна добрая женщина… остановилась над Лизою, лежавшею на земле, и старалась привести её в память*.

К пятой группе отнесен единственный оборот В ЦЕЛЯХ (!Род).

Как правило, это предлог: ● *Встреченные ими индейцы были настроены благожелательно, и Морган воспользовался этим* ***в целях*** *разведки.*

Редко (1,3% при выборке в 3400 предложений) это может быть сочетанием предлога и существительного: ● *Используется ли сертификат в целях, для которых он был выпущен?.*

Практически это словосочетание можно считать однозначным предлогом.

4. Заключение

Человек достаточно легко определяет, является ли данное словосочетание единым оборотом или распадается на отдельные слова. Очевидно, что при этом он опирается на смысл предложения. Но компьютер — не человек. Одна из возможных методик заключается в морфологическом анализе ближайшего контекста, а также в определении принадлежности слов ближайшего и удаленного контекста к определенным микрословарям. Наличие классификатора несколько расширяет эти возможности, но, по сути, классы — это ведь просто перечень большого количества слов, сгруппированных вокруг общего смысла. Анализ удаленного контекста достаточно сложен, но играет довольно большую роль особенно при разборе наречных оборотов — используется при разборе 25% словосочетаний. В состав удаленного контекста могут входить и некоторые знаки препинания. Приведем ряд примеров[3].

Вопрос в том, что эту девочку довели до истерики и ***до слез*** *за её мнение.*

Вот как-то (а именно: в прошлую субботу) довели училку, бедную… ***до слез****…*

*Вопрос был, как говорится, задан "****в лоб****".*

С юга дул слабый бриз прямо нам ***в лоб****, и небо было затянуто тучами.*

Радуевцы поняли, что ***в лоб*** *защитников им не взять.*

*Левитин нигде, кажется, "****в лоб****" самого этого слова не говорит.*

Убежала она без штанов ***в ночь*** *холодную.*

Я вышел на крыло мостика ***в ночь****.*

[3] В нижеприведенных примерах исследуемое словосочетание выделено жирным шрифтом.

*Запишите нам коллективное замечание **для формы**, а как человек — простите.*

*Поль тоже говорил потом, что если и было что-нибудь, то разве **для формы**, а в сущности они должны были сами знать, что опоздали.*

*Запрос, сформированный пользователем на РС, ожидает момента поступления необходимого сегмента **в память** данной РС.*

*То есть с младенческих лет люди приобщались к хору, к пению, к музыке, и эти звуки входили в их сознание, **в память** естественным путём.*

При разработке правил снятия омонимии в качестве обучающего корпуса бралось 50…70 предложений. Затем осуществлялся автоматический разбор всех предложений и проводилась экспертная оценка его правильности.

Так для словосочетания С КОРНЕМ было проанализировано 274 предложения из НКРЯ и дополнительного корпуса. Экспертная оценка — в этих предложениях 246 наречных оборотов. Результаты автоматического анализа: точность P = 0.961; полнота R = 0.996; F-мера F = 0.978.

Для словосочетания В СВЯЗИ С было проанализировано 368 предложений. Результаты: P = 1.000; R = 0.992; F = 0.996.

Отметим также, что в ряде случаев построенные нами правила автоматически позволяют снять семантическую омонимию. Типичным примером такого рода является правило для наречного оборота *с корнем* (см. выше). Если в каком-либо конкретном предложении это словосочетание является предлогом и существительным, то встает вопрос о классе последнего. Слово *корень* может относиться к разным семантическим классам: «Основа» (в этот класс входят, например, слова *база, базис, основа*), «Растения Части» (*береста, бутон, ветка*), «Математика Числа» (*делимое, дробь, множитель*) и «Слово Части_слова» (*морфема, префикс, суффикс*). В вышеприведенном правиле исследуемое словосочетание является предлогом с существительным в следующих случаях.

Если за словосочетанием следует существительное в родительном падеже единственного числа из класса «Растения» (например, *алтей, валериана, девясил*), то данной лексеме *корень* следует присвоить класс «Растения Части»;

если же за словосочетанием следует слово в кавычках, слово, оканчивающееся на дефис или написанное латинскими буквами, то данной лексеме *корень* следует присвоить класс «Слово Части_слова»;

Таким образом, в ряде случаев удается снять семантическую омонимию без дополнительного анализа.

Следует также отметить, что иногда однозначно решить вопрос о том, является ли данное словосочетание устойчивым оборотом или омонимичным ему свободным сочетанием слов практически невозможно, поэтому мы оцениваем правильность нашего разбора подобных словосочетаний на уровне 93–96%.

Литература

[1] Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. М.: МИЭМ, 2011.

[2] Национальный корпус русского языка // URL: http://www.ruscorpora.ru/ (дата обращения: 9.07.2013).
[3] *Кузнецов С.А.* Большой толковый словарь русского языка. СПб: Норинт, 1998.
[4] *Рогожникова Р.П.* Толковый словарь сочетаний, эквивалентных слову. М.: ООО «Издательство Астрель»: ООО «Издательство АСТ», 2003.
[5] Iomdin L., Petrochenkov V., Sizov V., Tsinman L. Etap parser: state of the art // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог». М.: РГГУ. Выпуск 11 (18). 2012. Том 2. С. 117 – 131.
[6] Anisimovich K.V., Druzhkin K.Ju., Minlos F.R., Petrova M.A., Selegey V.P., Zuev K.A. Syntactic and semantic parser based on ABBYY Compreno linguistic technologies // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог». М.:, РГГУ. Выпуск 11 (18). 2012. Том 2. С. 91 – 103.
[7] *Боярский К.К., Каневский Е.А.* Принцип работы семантико-синтаксического анализатора // Экономико-математические исследования: математические модели и информационные технологии. СПб: Нестор-История. 2012. С. 168 – 189.
[8] Боярский К.К., Каневский Е.А. Семантико-синтаксический парсер SEMSIN // Научно-технический вестник информационных технологий, механики и оптики. 2015. Т. 15, №5. С. 869 – 876.
[9] Каневский Е.А. Особые предложные обороты // Контрастивные исследования и прикладная лингвистика: матер. Междунар. науч. конф., (Минск, 29-30 октября 2014 г.). Минск: МГЛУ. 2015. Часть 1. С. 115 – 119.
[10] Боярский К.К., Каневский Е.А. Словосочетания, эквивалентные слову // Компьютерная лингвистика и вычислительная онтология: сборник научных статей. Труды XVIII объединенной научной конференции «Интернет и современное общество» (Санкт-Петербург, 23–25 июня 2015 г.) СПб: Университет ИТМО. 2015. С. 55 – 66.
[11] Каневский Е.А., Клименко Е.Н., Силина Е.Ф. Особые наречные обороты // Вторые чтения памяти профессора Б.Л. Овсиевича «Экономико-математические исследования: математические модели и информационные технологии»: Материалы Всероссийской конференции. (Санкт-Петербург, 26–28 октября 2015 г.). СПб: Нестор-История. 2015. С. 101 – 107.
[12] Тузов В.А. Компьютерная семантика русского языка. СПб: Изд-во С.-Петерб. ун-та, 2004.

Analysis of ambiguous collocations equivalent to the word

K. Boyarsky[1], E. Kanevsky[2], E. Klimenko[2], E. Silina[2]

[1] ITMO University

[2] St. Petersburg Institute for Economics and Mathematics RAS

The work is devoted to the analysis of collocations equivalent to a word. The ways of disambiguation at the collocations which on the one hand are the steady turns performing functions of adverbs and prepositions, and with another – representing a simple combination of a preposition with a noun are. It is argued that analysing the immediate context is often not enough to decide whether the collocations is a steady turn. Depending on the depth of use of the context, there are 4 groups of collocations. For disambiguation of expressions of the first group, it is sufficient to analyse the morphological features of the words in the nearest context, the second group requires checking for specific lemmas or word forms, the third – to analyse semantic classes of surrounding words, the fourth – to check a set of attributes for words of a remote context. Particular collocations having three options parsing considered separately: two of them represent different parts of speech and the third is a combination of a preposition with a noun. It is shown that, when the Russian text is automatically parsed, it is possible to select rules that make it possible to reliably remove homonymy from such collocations. Examples are given showing the significance of the remote context for the parsing.

Keywords: Sentence analysis, collocations, morphological analysis, parts of speech, lemmas, semantic classes.