Устойчивые словосочетания в роли предлогов

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

[1] Институт проблем региональной экономики РАН, [2] Университет ИТМО

Boyarin9@yandex.ru, eak300@mail.ru, alexklim2000@yandex.ru

Аннотация

В работе рассматриваются особенности обработки неоднословных оборотов (фразем), способных выполнять функции предлогов, при автоматическом анализе русскоязычных текстов. Сложность анализа таких оборотов заключается в достаточно высокой омонимичности, что снижает точность автоматического разбора предложений и, тем более, определение семантики. На основе возможных типов омонимии проведена классификация более 320 фразем.

К первой группе относятся фраземы, которые однозначно являются предлогами, но могут иметь семантическую неоднозначность. Для второй группы характерна частеречная омонимия предлог/наречие. К третьей группе отнесены фраземы, имеющие два/три варианта разбора: один/два оборота, соответствующие разным частям речи и простое сочетание предлога с существительным.

Для каждой группы приводятся списки наиболее частотных фразем (по НКРЯ). Указываются основы построения правил, позволяющих осуществлять эффективное снятие омонимии (применительно к парсеру SemSin). Приводятся примеры, показывающие, что анализ непосредственного окружения фраземы может быть недостаточен для снятия омонимии, так что необходимо рассмотрение удаленного контекста.

Ключевые слова: автоматический анализ текста, омонимия, словосочетания, лексические обороты, предложные группы

Библиографическая ссылка: Боярский К.К., Каневский Е.А., Клименко Е.Н. Устойчивые словосочетания в роли предлогов // Компьютерная лингвистика и вычислительные онтологии. Выпуск 5 (Труды XXIV Международной объединенной научной конференции «Интернет и современное общество», IMS-2021, Санкт-Петербург, 24 – 26 июня 2021 г. Сборник научных статей). — СПб.: Университет ИТМО, 2021. С. 17-28. DOI: 10.17586/2541-9781-2021-5-17-28

Введение

При автоматическом разборе предложений русского языка и построении дерева зависимостей возникает проблема снятия омонимии различных типов – морфологической, лексической, частеречной и т.д. Одним из путей ее решения является широкое использование стандартных сочетаний слов – фразем. Под этим термином понимают широкий спектр выражений с разной степенью идиоматичности [1]. Общим для них является то, что значение целого не является композицией значений составных частей. В общем случае слова, входящие в состав фразем, могут изменяться, однако в рамках данной работы нас интересуют неизменяемые фраземы, б*о*льшая часть которых является оборотами, выполняющими функции:

- наречий – в конце концов, время от времени, в свое время;
- предлогов – в зависимости от, в соответствии с, несмотря на;
- вводных оборотов – другими словами, к слову сказать;
- союзов – а также, в том числе и, вместо того чтобы, если бы;

– частиц – все же, как бы, как раз, к тому же, едва не, вроде бы;
– предикативных оборотов – не дай бог, лыка не вяжет.

Наиболее полные списки оборотов приведены в НКРЯ [2]. Также нами использовались словари С.А. Кузнецова [3] и Р.П. Рогожниковой [4].

В настоящее время пристальное внимание привлекает семантика предложных групп, в том числе тех, где в качестве предлога выступают неоднословные сочетания [5]. Как показывает даже предварительный анализ, большинство таких фразем не обладают омонимией и всегда являются предложными оборотами.

Однако возможна ситуация, когда одно и то же сочетание нескольких слов может соответствовать двум различным оборотам. Так, например, фразема *без сопровождения* может выполнять функции или предлога, или наречия в зависимости от наличия контекста справа: слова в родительном падеже, глагола или знака препинания:

● *Сегодня первый день идём **без сопровождения** чаек, они отстали.*

● *Путешествовал львенок из Петербурга в Москву один, **без сопровождения**, в небольшой клетке*[1].

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. Так, например, фразема *под знаком* может выполнять функции или предлога, или остаться свободным словосочетанием в зависимости от наличия справа слова в родительном падеже:

● *Первые шесть месяцев 2012-го года прошли **под знаком** поиска новых выставочных пространств.*

● *Припарковать машину оказалось негде, ее оставили на единственном свободном пятачке, как потом выяснилось, под знаком, запрещающим парковку.*

Все предложные обороты (и соответствующие фраземы) на наш взгляд, можно разделить на три группы, которые мы и рассмотрим ниже.

Известны два подхода к разбору подобных оборотов. Первый подход не предполагает какого-либо их специального графематического их выделения – примером является парсер «Этап-3» [6]. При втором подходе такой оборот выделяется особым образом – примером является парсер фирмы ABBYY [7]. В последнее время эти подходы сближаются, и в последней версии ЭТАП-4 [8] часть оборотов тоже выделяется. Принцип работы нашего парсера SemSin [9], с помощью которого анализируются предложные обороты, близок ко второму варианту. Неднословные фраземы объединяются в один токен [10].

В состав парсера SemSin входят 4 блока: словарь, морфологический анализатор, продукционные правила и лексический анализатор. Очередной абзац русскоязычного текста подвергается морфологическому анализу с выделением отдельных токенов (слов, словосочетаний, знаков препинания, чисел и т. д.). Затем цепочка токенов обрабатывается в лексическом анализаторе с помощью системы продукционных правил, целью которых является преобразование линейной последовательности токенов в дерево зависимостей.

Принципы построения словаря парсера основаны на идеях В.А. Тузова [11] Основная таблица словаря содержит более 195 тыс. лексем, распределенных по 1700 классам [12]. Каждая лексема имеет морфологические характеристики, а также номер своего семантического класса и актанты или валентности (для подключения зависимых слов) в виде падежей (!Им, !Род, !Вин и т. д.) или предлогов с соответствующими падежами (!вВин, !наПред и т. д.). Часто перед таким актантом указаны допустимые классы слов, могущих их замещать. Около 14% слов в словаре имеют две и более лексемы.

[1] Здесь и далее все примеры взяты из НКРЯ и отделены знаком «●», жирным шрифтом в примерах выделены фраземы, являющиеся оборотами. Подчеркнуты слова, которые позволяют принять то или иное решение.

Помимо основной таблицы, имеются вспомогательные, которые и обеспечивают выполнение задач, представляющих интерес в рамках данной работы. Это таблица словосочетаний (более 5350 строк), содержащая устойчивые сочетания слов с разными типами словоизменения. Это могут быть коллокации (*вид на жительство*), названия организаций (*Чейз Манхеттен Банк*) или идиоматические выражения (*белая ворона*). В этих случаях одно или все слова могут употребляться в разных словоформах.

В данной работе нас интересуют неизменяемые фраземы, которые образуют составные предлоги, наречия и т.д. Если парсер принимает решение, что некоторое словосочетание является такой фраземой, то входящие в него слова объединяются в один токен, которому приписываются соответствующие параметры управления (возможные падежи и классы подключаемых слов) из словаря фразем.

Второй вспомогательной таблицей является таблица предлогов (более 2460 строк) с указанием падежей и семантических классов подключаемых существительных. Если связь предлога со слугой носит синтаксический характер и, как правило, совпадает с падежом слуги, то подключение предложной группы к хозяину больше отражает семантику (Где, Когда, Почему и т.д.).

Группа 1. Предложные обороты без лексической омонимии

Переходя к анализу предложных оборотов, отметим, что из трех групп самой большой является первая группа, обороты которой не имеют омонимов и являются однозначными[2]. Группа состоит из двух подгрупп: фраземы первой оканчиваются существительными (1А), второй – предлогами (1Б).

Подгруппа 1А

К этой подгруппе относятся однозначные предложные обороты, фраземы которых состоят из двух слов: предлога и существительного. В нашем словаре насчитывается более 60 таких оборотов.

Подавляющее большинство из них требует после себя родительного падежа.

Например: ***В целях*** *профилактики при чтении документов с Crypto полезно заблокировать запись механическим микропереключателем*.

Другие обороты требуют после себя дательного падежа: *в противовес, в противоположность, в ущерб, на благо, на радость*.

Например: *Отдавать все силы организации выборов* ***в ущерб*** *профессиональной деятельности*.

Большинство предложных оборотов этого типа могут связываться с хозяином только одной связью. Однако имеются и такие обороты с семантической омонимией, которые для подключения к хозяину имеют две связи, выбор одной из которых зависит от хозяина (его класса, его внутренних актантов или вообще от его части речи). Таковы следующие обороты: *в честь* (Зачем, Почему), *из числа* (Изо, Какой), *на основе* (Как, Какой), *по поводу* (поДат, Почему).

Семантика предложных связей подробно рассмотрена в словаре Г.А. Золотовой [13], однако там отсутствуют формальные правила, позволяющие соотнести преимущественно синтаксические связи, вырабатываемые парсером, с семантикой из [13]. Это достаточно сложная задача, которая решается пока в некоторых частных случаях [14].

Ниже приведены примеры с двумя предложными оборотами, причем связь с хозяином предлога представлена в скобках в терминах парсера SemSin и в семантических связях Золотовой.

---

[2] Русский язык характерен большим количеством исключений. Здесь и далее мы считаем допустимым, если количество ошибок не превосходит 3-5%.

● *В прессе отмечалось, что это был салют в 101 залп* ***в честь*** *возникшего в России рабочего вопроса* (был – Зачем (финитив) – в честь).

● *И вдруг Олег вспомнил, как однажды он был на торжественном ужине, устроенном* ***в честь*** *приезда английского принца Чарльза* (устроенном – Почему (каузатив) – в честь).

● *Если опальный магнат будет исключён* ***из числа*** *сопредседателей ЛР, у партии возникнут финансовые затруднения, полагают влиятельные эксперты* (исключён– Изо (финитивно-фазисное) – из числа).

● *Обычно в то время, как наверху происходила церемония награждения победителей, внизу совершалась казнь изменников, трусов, неудачников* ***из числа*** *подданных Великого курфюрста* (неудачников– Какой (генеративный) – из числа).

В таблице 1 приведены наиболее часто встречающиеся предложные обороты этой подгруппы, требующие после себя родительного падежа.

Таблица 1. Неизменяемые предложные обороты

Оборот	Требуемый падеж	Связь с хозяином	Встречаемость
В ВИДЕ	Род	Как	11419
В ГЛУБЬ	Род	Куда	2701
В ПОЛЬЗУ	Род	Как	4899
В ПРИСУТСТВИИ	Род	Как	6605
В СТОРОНУ	Род	Куда	33598
В ТЕЧЕНИЕ	Род	какДолго	18613
В ХОДЕ	Род	Когда	5991
В ЦЕЛЯХ	Род	Для	2648
В ЧЕСТЬ	Род	Зачем, Почему	2540
ВО ВРЕМЯ	Род	Когда	68458
ВО ИМЯ	Род	Зачем	4200
ДЛЯ СОЗДАНИЯ	Род	Зачем	2193
ЗА ИСКЛЮЧЕНИЕМ	Род	Как	3842
ЗА ПРЕДЕЛЫ	Род	Куда	3952
ЗА СЧЕТ	Род	Как	8775
ИЗ ЧИСЛА	Род	Изо, Какой	3810
НА ОСНОВАНИИ	Род	Почему	10224
НА ОСНОВЕ	Род	Как, Какой	7547
НА ПРОТЯЖЕНИИ	Род	какДолго	3911
ПО ПОВОДУ	Род	поДат, Почему	9757

Подгруппа 1Б

К этой подгруппе относятся однозначные предложные обороты, фраземы которых состоят из двух, трех или четырех слов и оканчиваются предлогом. В нашем словаре насчитывается 153 таких оборота. Практически все предложные обороты, оканчивающиеся предлогом, относятся к этой подгруппе. Нам известно только три исключения: фраземы *на глазах у*, *под носом у* и *под самым носом у*. Действительно, сравним два предложения: *Ты* ***на глазах у*** *зрителя вершишь свой путь* и *Стали мы во дворе, и вижу я:* ***на глазах у*** *него будто слеза поблескивает*. Совершенно очевидно, что в первом из них фразема является предложным оборотом, а во втором – просто свободным сочетанием трех слов. Аналогично обстоит ситуация и с двумя другими фраземами. Все они относятся к третьей группе.

Большинство таких оборотов начинается с предлога, чаще всего это «в». В конце чаще всего расположены предлоги «с» «со» или «от». Требуемый после оборота падеж определяется концевым предлогом, хотя возможны и другие варианты. Примеры наиболее частотных оборотов этой группы приведены в таблице 2.

Большинство предложных оборотов этого типа могут связываться с хозяином только одной связью. Однако имеются и такие обороты, которые для подключения к хозяину имеют несколько связей, выбор одной из которых зависит от хозяина (его класса, его внутренних актантов или вообще от его части речи). Таковы следующие обороты: *верхом на* (Как, Куда), *вплоть до* (Как, доКогда, Докуда, Сколько), *начиная от* (Как, Когда), *начиная с* (Как, Когда), *начиная со* (Как, Когда), *совместно с* (Как, сТв), *совместно со* (Как, сТв).

Таблица 2. Фраземы, оканчивающиеся предлогом

Оборот	Требуемый падеж	Связь с хозяином	Встречаемость
В ОДНОЙ ИЗ	Род	Где	5722
В ОДНОМ ИЗ	Род	Где	9257
В ОТВЕТ НА	Вин	Как	5820
В ОТЛИЧИЕ ОТ	Род	Как	6901
В СВЯЗИ С	Тв	Почему	11357
В СООТВЕТСТВИИ С	Тв	Как	10184
ВМЕСТЕ С	Тв	Как	28944
ВМЕСТЕ СО	Тв	Как	8731
ВНЕ ЗАВИСИМОСТИ ОТ	Род	Как	10335
ВПЛОТЬ ДО	Род	Как, доКогда, Докуда, Сколько	8897
ВСЛЕД ЗА	Тв	Как	6885
НЕ БЕЗ	Род	сТв	12256
НЕ ДО	Род	До	10337
НЕСМОТРЯ НА	Вин	Как	14671
ПО НАПРАВЛЕНИЮ К	Дат	Куда	27789
ПО ОТНОШЕНИЮ К	Дат	поОтн	13152
ПО СРАВНЕНИЮ С	Тв	Как	6546
РЯДОМ С	Тв	Где	17881
ЧТО ДО	Род	Как	7174

Ниже приведены примеры для предложного оборота *вплоть до*.

Помощь готова оказать любую, вплоть до аврального написания сочинения (оказать – Как (Интенсив) – вплоть до).

Но за первые 24 года, то есть вплоть до 1924 г. включительно, лишь дважды премию получал не один лауреат (получал – доКогда (Темпоратив) – вплоть до).

На снимки попадало как само Солнце, так и его окрестности вплоть до Венеры (попадало – Докуда (Директив) – вплоть до).

С помощью частиц, разогнанных на ускорителях, мы можем сегодня зондировать расстояния вплоть до 10-16 (зондировать– Сколько (Дименсив-квантитатив) – вплоть до).

Как указывалось выше, при синтаксическом разборе предложения в парсере SemSin неизменяемый фразеологизм рассматривается как один токен. Предсинтаксический модуль, просматривая справа налево поступающий текст и сравнивая его со строками таблицы фразем, выделяет очередную фразему. При этом просмотр справа налево обеспечивает достаточно простой анализ таких ситуаций когда несколько предложных оборотов отличаются только последним словом (например, в согласии, в согласии с, в согласии со).

Группа 2. Обороты с омонимией предлог/наречие.

К этой группе относятся самые простые омонимичные предложные обороты, фраземы которых могут выполнять функции предлогов и наречий [15].

В нашем словаре насчитывается 24 таких оборота. Например, фразема *в конце* может быть предлогом, если после нее находится слово в родительном падеже или наречием при его отсутствии:

Принятая в конце января с. г. резолюция 1526 поставила перед Комитетом целый ряд новых задач…

Только в конце обратил внимание, что при сражении на мечах, они ими по большей части над головами махали.

Подавляющее большинство предложных оборотов требует после себя родительного падежа. Например:

Родиться князем не мудрено, и можно по праву породы называться сиятельством. Два оборота требуют после себя дательного падежа: в угоду, не в пример. Например:

Он просто не хотел никого казнить в угоду иудеям.

Большинство предложных оборотов этого типа могут связываться с хозяином только одной связью. Однако наблюдается несколько оборотов, которые для подключения к хозяину имеют две связи, выбор одной из которых зависит от хозяина (его класса, его внутренних актантов или вообще от его части речи). Таковы следующие обороты: *в конце* (Где, Когда), *в начале* (Где, Когда), *в середине* (Где, Когда), *к концу* (Когда, Куда), *к началу* (Когда, Куда). Ниже приведены примеры для предложного оборота *в начале*.

Я только успел заметить далеко **в начале** улицы две светлых фигурки (заметить – Где (Локатив) – в начале).

Да, а **в начале** марта мы-таки устроим массовый вылет (устроим – Когда (Темпоратив) – в начале).

В таблице 3 приведены наиболее часто встречающиеся предложные обороты второй группы.

Таблица 3. Наиболее частотные обороты второй группы

Оборот	Требуемый падеж	Связь с хозяином	Встречаемость
В КОНЦЕ	Род	Где, Когда	52203
В НАЧАЛЕ	Род	Где, Когда	26718
В ПОДТВЕРЖДЕНИЕ	Род	Как	1129
В РАМКАХ	Род	Как	10169
В СЕРЕДИНЕ	Род	Где, Когда	9474
ВО ГЛАВЕ	Род	Где	13013
К КОНЦУ	Род	Когда, Куда	11569
К НАЧАЛУ	Род	Когда, Куда	3600
НА КРАЮ	Род	Где	4386
НЕ СЧИТАЯ	Род	Как	2483
ПО АДРЕСУ	Род	Как	3406
ПО ПОРУЧЕНИЮ	Род	Почему	1550
ПО ПРАВУ	Род	Почему	2284
ПО ПРОСЬБЕ	Род	Почему	2431
ПО СЛУЧАЮ	Род	Почему	8081
СО СТОРОНЫ	Род	Откуда	29189

При обнаружении предсинтаксическим модулем фразем второй группы выдается две лексемы: предлог и наречие. Дальше запускается правило под название «Предлог-Наречие», которое и осуществляет окончательный выбор. Поскольку это правило срабатывает после формирования именной группы, то проверка падежа производится у центра именной группы, что и обеспечивает правильный выбор из этих двух лексем. Например, в следующем предложении фразема играет роль предлога: *Они образуются даже **в рамках** уже действующих транспортных структур*.

Группа 3. Словосочетания, которые могут не быть фраземами

К этой группе относятся сложные омонимичные предложные обороты, фраземы которых могут выполнять функции предлогов или являться простым сочетанием слов. В первом случае все слова, составляющие фразему, должны быть объединены в один токен, во втором – должны быть оставлены без изменения. Таким образом, предсинтаксический модуль, выделив очередную фразему, относящуюся к третьей группе, сам не может объединять ее токены в один. Для дальнейшей обработки фраземы запускается практически первое по порядку правило, которое и принимает решение о том, будет ли эта фразема предложным оборотом или нет. В нашем словаре имеется 87 таких оборотов.

Наиболее подробно такого рода словосочетания рассмотрены в книге Р.П. Рогожниковой [4], которая анализирует их с семантической точки зрения. Однако для компьютерной лингвистики на сегодня это недоступно, так что нам приходится учитывать только окружающий контекст, его грамматику и классы. Иногда приходится учитывать и удаленный контекст.

В связи с таким подходом можно разделить предложные обороты этой группы на 3 подгруппы, в зависимости от сложности их анализа.

Подгруппа 3А

К этой подгруппе относятся омонимичные фраземы, которые могут выполнять функции предлогов в случае выполнения самого простого условия. Этим условием является наличие справа слова в родительном падеже. При отсутствии такого падежа фразема остается простым сочетанием слов. Например:

Проверки бизнеса должны проводиться ***с ведома*** *прокуратуры*.

Застройка в СНТ велась с ведома и при попустительстве местных властей.

В состав этой подгруппы входит 13 оборотов, наиболее часто встречающиеся приведены в таблице 4. Как и ранее встречаются предложные обороты, которые могут подключаться к хозяину различными связями. Например:

Такие счета могут быть номинированы в иностранной валюте, а владельцы счёта NRI могут определять бенефициария ***в пределах*** *Индии* (определить – Где (Директив)– *в пределах*).

Отступления сделаны для пироксенов, гранатов, хлоритов и амфиболов, поскольку минералы ***в пределах*** *этих групп близки по условиям формирования...* (*близки* – Как (Характеристика способа или меры действия) – *в пределах*).

Таблица 4. Наиболее частотные обороты подгруппы 3А

Оборот	Требуемый падеж	Связь с хозяином	Встречаемость
В ПРЕДЕЛАХ	Род	Где, Как	7666
В СЛУЧАЕ	Род	Когда	22967
В СФЕРЕ	Род	Как	4969
В ЧИСЛЕ	Род	вПред	10380
В ЧИСЛО	Род	вВин	2448
С ЦЕЛЬЮ	Род	Зачем	9717

Подгруппа 3Б

К этой подгруппе относятся омонимичные фраземы, которые могут выполнять функции предлогов или остаться простым сочетанием слов. Для выбора того или иного варианта требуется выполнить сложное условие. Для его реализации нам приходится учитывать окружающий контекст, грамматику и классы отдельных слов. Иногда приходится учитывать даже и удаленный контекст в пределах всего предложения [16].

Например, фразема *на глазах у* может иметь два семантических значения: что-то происходит с глазами кого-то (и это будет свободным сочетанием трех слов) или что-то происходит в присутствии кого-то (и это будет предложный оборот). Для анализа такой фраземы следует исполнить следующее правило: если слева или справа от фраземы в пределах семи слов встречается одно из слов *слезы, слеза, влага*, то это простое словосочетание, в противном случае это предложный оборот. Заметим, что в обоих случаях после фраземы стоит слово в родительном падеже:

На глазах у Маруси появились слезы.

***На глазах у** посетителей, так и не слезших со столов, ему удалось поймать 28 змей.*

В состав этой подгруппы входит 57 оборотов, наиболее часто встречающиеся приведены в таблице 5. Как и ранее встречаются предложные обороты, которые могут подключаться к хозяину различными связями. Например: *Заседание Госдумы **по вопросу** его ратификации состоится 20 или 21 марта* (*Заседание* – Какой – *по вопросу*). ***По вопросу** губернатора Резанов догадался, что тот значительно больше его осведомлен* (*догадался* – Про – *По вопросу*). *Самым ярым оппонентом Кука **по вопросу** распространения американских культурных растений в области Тихого океана много лет был его соотечественник Меррилл* (*оппонентом* – поДат – *по вопросу*).

Таблица 5. Наиболее частотные обороты подгруппы 3Б

Оборот	Требуемый падеж	Связь с хозяином	Встречаемость
В ГЛАЗАХ	Род	вПред	16236
В КАЧЕСТВЕ	Род	Как	32100
В ОБЛАСТИ	Род	вПред	10024
В ОТНОШЕНИИ	Род	вПред	15663
В ПОРЯДКЕ	Род	Как	13555
В РАЙОНЕ	Род	Где	10941
В СИЛУ	Род	Как	13417
С ПОМОЩЬЮ	Род	Как	22032
С ТОЧКИ ЗРЕНИЯ	Род	Как	10843

Подгруппа 3В

К этой подгруппе относятся самые сложные омонимичные фраземы, которые могут выполнять функции предлогов, наречий или остаться простым сочетанием слов. Для выбора того или иного варианта требуется выполнить достаточно громоздкое условие. В общем случае для его реализации нам приходится учитывать окружающий контекст, грамматику и классы отдельных слов. Иногда приходится учитывать даже и удаленный контекст в пределах всего предложения. Рассмотрим, например, фразему *в результате*. В книге Рогожниковой для нее приведено некоторое семантическое обоснование и примеры [4]. Основываясь на этом, для ее анализа нами было разработано следующее правило.

Если слева от фраземы стоят леммы СОМНЕВАТЬСЯ, СОМНЕНИЕ, УВЕРЕННЫЙ, а справа (непосредственно или через одно слово в родительном падеже) расположены леммы АНАЛИЗ, ГОЛОСОВАНИЕ, ИССЛЕДОВАНИЕ, ОПЕРАЦИЯ, ОПЫТ, ТЕСТ, ЭКСПЕРИМЕНТ, словоформы которых имеют родительный падеж, то в этом случае фразема является простым сочетанием слов. Например:

Не будучи уверен в результате голосования и не желая идти на риск и в то же время сильно надеясь на воздействие ленинской речи, левый блок сделал уступку…

Я не сомневался в результате этого эксперимента.

Если слева от фраземы стоят леммы СОМНЕВАТЬСЯ, СОМНЕНИЕ, УВЕРЕННЫЙ, а справа – запятая или точка, то фразема также является простым сочетанием слов:

Мой добрый друг был, как правило, уверен в результате.

Если справа от фраземы стоит слово в родительном падеже, то она выполняет функцию предлога:

*Это сообщение выдаётся автоматизированной системой, если **в результате** вычисления формула получила значение "ложь".*

В противном случае фразема выполняет функцию наречия:

***В результате** объекты имитационной модели перейдут в некорректные состояния.*

Таким образом видно, что формализовать семантические отношения, в принципе, можно, но иногда это выливается в достаточно громоздкие правила.

В состав этой подгруппы входит 15 оборотов, наиболее часто встречающиеся приведены в таблице 6. Следует отметить, что, по крайней мере, две фраземы из них имеют и более трех вариантов омонимии. Так фраземы *в меру* дополнительно может быть предикатом: *Вроде бы все **в меру**, все на своих местах.* Фразема *в разрезе* дополнительно может выполнять функцию определения: *У меня над кроватью, сколько себя помню, висел план огромного океанского парохода **в разрезе**.*

Таблица 6. Наиболее частотные обороты подгруппы 3Б

Оборот	Требуемый падеж	Связь с хозяином	Встречаемость
В ЗАКЛЮЧЕНИЕ	Род	Где,Когда	4740
В МЕРУ	Род	Как	3071
В РЕЗУЛЬТАТЕ	Род	Как	26141
ЗА РАМКИ	Род	Куда	1209
НА РАССТОЯНИИ	Род	Где	4111
НА СТОРОНЕ	Род	Где	3694
НА ФОНЕ	Род	Как	7763
ПО ОКОНЧАНИИ	Род	Когда	6262
ПО ПУТИ	Род	Куда	6038

Заключение

В результате проведенного исследования выполнена классификация устойчивых оборотов (фразем) в зависимости от типа омонимии. Разработаны правила, позволяющие с высокой точностью снимать частеречную и синтаксическую омонимию. Мы полагаем, что в связи с большой вариативностью русского языка повышение точности разбора ряда конструкций выше 95% может требовать непропорционально больших усилий, и, фактически, сводится к анализу конкретных фраз. Поэтому в некоторых случаях редко встречающиеся варианты игнорировались. Например, конструкция *под знаком + род. пад.* встречается в основном и газетном корпусах НКРЯ свыше 1700 раз, при этом только в 9 случаях мы обнаружили, что это свободное словосочетание, а не составной предлог (*под знаком интеграла…*).

В то же время снятие семантической омонимии представляет собой значительно более сложную задачу, требующую дополнительных исследований.

Литература

[1] Копотев М,В,, Стексова Т,И, Исключение как правило: Переходные единицы в грамматике и словаре. М.: Языки славянской культуры: Рукописные памятники Древней Руси, 2016.

[2] Национальный корпус русского языка. – URL: http://www.ruscorpora.ru/ (дата обращения: 22.02.2021).

[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», 2003.
[5] Zakharov V., Golovina A., Alexeeva E., Gudkov V. Russian Secondary Prepositions: Methodology of Analysis //XVI Международная конференция по компьютерной и когнитивной лингвистике (TEL 2020) (в печати).
[6] 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. Т. 2. С. 117–131.
[7] 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.
[8] Лингвистический процессор ЭТАП-4. – URL: http://proling.iitp.ru/ru/etap4 (дата обращения: 22.02.2021).
[9] 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.
[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» (IMS-2015), Санкт-Петербург, 23–25 июня 2015 г. – СПб.: Университет ИТМО, 2015. С. 55–66.
[11] Тузов В.А. Компьютерная семантика русского языка. СПб.: Изд-во С.-Петерб. ун-та, 2004.
[12] Боярский К.К., Каневский Е.А, Стафеев С.К.. Использование словарной информации при анализе текста // Научно-технический вестник информационных технологий, механики и оптики. 2012. №3(79). С. 87–91.
[13] Золотова Г.А. Синтаксический словарь. М.: Едиториал УРСС, 2011. – 440 с.
[14] Zakharov V., Boyarsky R., Golovina A., Kozlova A. Semantic Analysis of Russian Prepositional Constructions // RASLAN 2020. Recent Advances in Slavonic Natural Language Processing. Proceedings. Brno, 2020. P. 103–113.
[15] 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 26–28 октября 2015 года. – СПб.: Нестор-История, 2015. С. 101–107.
[16] Каневский Е.А. Особые предложные обороты // Контрастивные исследования и прикладная лингвистика: матер. Междунар. науч. конф., Минск, 29-30 октября 2014. Часть 1. – Минск: МГЛУ, 2015. С. 115–119.

Stable Phrases as Prepositions

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

[1] Institute of Regional Economics Problems RAS, [2] ITMO University

The paper discusses the features of processing multiword phrases (phrasemes) that can play the role of prepositions in the automatic analysis of Russian-language texts. The complexity of the analysis of such phrases lies in a sufficiently high homonymy, which reduces the accuracy of automatic parsing of sentences and, moreover, the definition of semantics. On the basis of possible types of homonymy, more than 320 phrasemes have been classified.

The first group includes phrasemes that can unambiguously be recognized as prepositions but may have semantic ambiguity. The second group is characterized by part-of-speech homonymy of preposition / adverb. The third group includes phrasemes that have two / three parsing options: one / two phrases corresponding to different parts of speech and a simple combination of a preposition with a noun. For each group, lists of the most frequent phrasemes (according to RNC) are given. The basics of building rules that allow an effective removal of homonymy (as applied to the SemSin parser) are indicated. Examples are given to show that the analysis of the immediate environment of the phraseme may not be sufficient to remove the homonymy, but it is necessary to consider the remote context.

Keywords: automatic text analysis, homonymy, lexical phrases, prepositional groups

Reference for citation: Boyarsky K., Kanevsky E., Klimenko E. Stable Phrases as Prepositions // Computer Linguistics and Computing Ontologies. Vol. 5 (Proceedings of the XXIV International Joint Scientific Conference «Internet and Modern Society», IMS-2021, St. Petersburg, June 24-26, 2021). - St. Petersburg: ITMO University, 2021. P. 17 – 28. DOI: 10.17586/2541-9781-2021-5-17-28

Reference

[1] Kopotev M.V., Steksova T.I., Isklyuchenie kak pravilo: Perekhodnye edinicy v grammatike i slovare. — M.: Yazyki slavyanskoj kul'tury: Rukopisnye pamyatniki Drevnej Rusi, 2016. (In Russian).
[2] Natsional'nyy korpus russkogo yazyka [Elektronnyy resurs] // URL: http://www.ruscorpora.ru/ (data obrascheniya: 19.02.2021). (In Russian).
[3] Kuznetsov S.A. Bol'shoy tolkoviy slovar russkogo yazika. SPb.: Norint, 1998. (In Russian).
[4] Rogozhnikova R.P. Tolkovyj slovar' sochetanij, ekvivalentnyh slovu. M.: OOO «Izdatel'stvo Astrel'», 2003. (In Russian).
[5] Zakharov V., Golovina A., Alexeeva E., Gudkov V. Russian Secondary Prepositions: Methodology of Analysis //XVI Mezhdunarodnaya konferenciya po komp'yuternoj i kognitivnoj lingvistike (TEL 2020).
[6] IomdinL., Petrochenkov V., Sizov V., Tsinman L. Etap parser: state of the art // Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Po materialam ezhegodnoj Mezhdunarodnoj konferencii «Dialog». M., RGGU, vypusk 11 (18), 2012. V. 2. P. 117–131.
[7] 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 // Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Po materialam ezhegodnoj Mezhdunarodnoj konferencii «Dialog». M:, RGGU, vypusk 11 (18), 2012. V. 2. P. 91–103.
[8] Lingvisticheskiy protsessor ETAP-4 [Elektronnyy resurs] // URL: http://www. http://proling.iitp.ru/ru/etap4 (data obrascheniya: 22.02.2021). (In Russian).
[9] Boyarsky K.K., Kanevsky E.A. Semantiko-sintaksicheskiy parser SEMSIN // Nauchno-tekhnicheskiy vestnik informatsionnyh tekhnologiy, mekhaniki i optiki. 2015, V. 15, №5. – P. 869–876. (In Russian).
[10] Boyarsky K.K., Kanevsky E.A. Slovosochetaniya, ekvivalentnye slovu // Komp'yuternaya lingvistika i vychislitel'naya ontologiya: sbornik nauchnyh statej. Trudy XVIII ob"edinennoj nauchnoj konferencii «Internet i sovremennoe obshchestvo» (IMS-2015) – SPb: ITMO University, 2015. P. 55–66. (In Russian).
[11] Tuzov V.A. Komp'yuternaya semantika russkogo yazyka. SPb. Izd-vo S.-Peterb. un-ta, 2004. (In Russian).
[12] Boyarsky K.K., Kanevsky E.A., Stafeev S.K. Ispol'zovanie slovarnoj informacii pri analize teksta // Nauchno-tekhnicheskij vestnik informacionnyh tekhnologij, mekhaniki i optiki. 2012-№3(79). P. 87–91. (In Russian).
[13] Zolotova G.A. Sintaksicheskij slovar'. M.: Editorial URSS, 2011. 440 p. (In Russian).

[14] Zakharov V., Boyarsky R., Golovina A., Kozlova A. Semantic Analysis of Russian Prepositional Constructions // RASLAN 2020. Recent Advances in Slavonic Natural Language Processing. Proceedings. Brno, 2020. P. 103–113.
[15] Kanevskij E.A., Klimenko E.N., Silina E.F. Osobye narechnye oboroty // Vtorye chteniya pamyati professora B.L. Ovsievicha «Ekonomiko-matematicheskie issledovaniya: matematicheskie modeli i informacionnye tekhnologii»: Materialy Vserossijskoj konferencii. – SPb.: Nestor-Istoriya, 2015. P. 101–107. (In Russian).
[16] Kanevsky E.A. Osobye predlozhnye oboroty // Kontrastivnye issledovaniya i prikladnaya lingvistika: mater. Mezhdunar. nauch. konf., Minsk, 2014. Part 1. Minsk: MGLU, 2015. P. 115–119. (In Russian).