УДК 004.912;81'322.3

Волкова И.А., Головин И.Г.

ЛИНГВИСТИЧЕСКИЙ ПРОЦЕССОР РУССКОГО ЯЗЫКА: АНАЛИЗ УСТОЙЧИВЫХ СЛОВОСОЧЕТАНИЙ

Московский государственный университет им. М.В.Ломоносова,
факультет вычислительной математики и кибернетики

Volkova I.A., Golovin I.G.

RUSSIAN LINGUISTIC PROCESSOR: PARSING OF STABLE PHRASEOLOGICAL EXPRESSIONS

Аннотация. В работе рассматривается специфика автоматического анализа устойчивых словосочетаний, структура словаря устойчивых словосочетаний русского языка, а также метод синтаксического анализа фраз русского языка с учетом устойчивых словосочетаний.

Ключевые слова: устойчивое словосочетание, фразеологизм, сетевая грамматика, предсинтаксический анализ, синтаксический анализ.

Abstract. We describe the specific features of automatic parsing of stable phraseological expressions and the structure of Russian language stable expressions dictionary. Also the method of Russian phrase parsing using is described which considers the stable expressions.

Key words: stable phraseological expression, network grammar, syntactic analysis, presyntactic analysis.

Вступление.

В настоящее время актуальной является задача построения различных систем автоматической обработки текстов (АОТ-систем), основу которых составляют

лингвистические процессоры – программные комплексы, в той или иной мере осуществляющие анализ и/или синтез текстов на естественном языке. Традиционно в лингвистических процессорах в отдельные подсистемы выделяются программы, работающие на разных языковых уровнях: морфологическом, синтаксическом и семантико-прагматическом [1].

Задача разработки морфологических компонентов лингвистических процессоров на сегодняшний день практически решена.

Существующие семантико-прагматические компоненты в основном носят экспериментальный характер или рассчитаны на работу в узких, ограниченных проблемных областях.

Задачу разработки синтаксических компонентов пока нельзя считать решенной, поскольку существующие синтаксические анализаторы наряду с достоинствами, обладают и рядом недостатков: обладают не всегда приемлемой точностью анализа, не всегда адекватно решают проблему синтаксической омонимии, накладывают более или менее серьезные ограничения на анализируемые синтаксические конструкции.

Данная работа посвящена проблеме совершенствования качества синтаксического анализа русскоязычных текстов за счет выявления и анализа устойчивых словосочетаний русского языка, в частности, фразеологизмов.

Обзор литературы.

Наиболее известными и развитыми на сегодняшний день можно назвать следующие синтаксические анализаторы русского языка: Link Grammar Parser [2], синтаксические анализаторы систем ЭТАП3 [3], COMPRENO [4], Cognitive Dwarf [5].

Перечисленные синтаксические анализаторы различаются используемыми обозначениями, способом представления синтаксической структуры, классификацией типов связей. В некоторых системах отличаются решения вопроса о направлении синтаксической зависимости между двумя словами

(предлог – существительное, связи в сочинительных конструкциях и прочие) и некоторые другие.

Все эти анализаторы имеют свои достоинства и недостатки. Каждый из них справляется с простыми синтаксическими конструкциями, но практически все они получают неправильные результирующие синтаксические структуры в случае сложных входных данных или вообще отказываются их обрабатывать.

На данный момент многие синтаксические анализаторы не работают с такими сложными языковыми конструкциями, как фразеологизмы, междометия, обращения, вводные конструкции, сложные числительные, списочные конструкции и некоторые другие.

В данной работе рассматривается метод выявления и анализа устойчивых словосочетаний русского языка с целью повышения качества синтаксического анализа.

Поскольку устойчивые словосочетания (фразеологизмы, научные термины и сложные имена собственные) зачастую играют особую синтаксическую роль в предложениях, данная задача требует отдельного решения. К примеру, в предложении *бежал во всю прыть* фразеологизм означает не дополнение (бежал куда?), а обстоятельство действия (бежал как?), т.е. играет роль не именной группы, а наречной. Однако современные анализаторы не всегда адекватно обрабатывают такие структуры.

Кроме того, учет устойчивых словосочетаний на синтаксическом уровне позволит на этапе семантического анализа правильно определить смысл входной фразы, который в случае фразеологизмов не определяется из смысла входящих в него слов.

Входные данные и методы.

Задачу анализа устойчивых словосочетаний можно разбить на следующие подзадачи:

- классификация устойчивых словосочетаний;

- разработка структуры словарной статьи и создание словаря устойчивых словосочетаний;
- разработка и реализация метода предсинтаксического анализа для выявления устойчивых словосочетаний;
- построение расширенной сети переходов грамматики русского языка и реализация по ней алгоритма синтаксического анализа с учетом устойчивых словосочетаний.

Классификация устойчивых словосочетаний

По результатам анализа различных фразеологизмов [6,7] рассматриваются следующие типы словосочетаний: разделимые, внутрь которых могут быть добавлены уточняющие слова, и неразделимые, изменяемые и неизменяемые. Среди изменяемых словосочетаний выделяют шесть видов словосочетаний в зависимости от типа и количества изменяемых слов.

Разделимость словосочетаний

В первую очередь свойство разделимости присуще фразеологизмам. Различают разделимые и неразделимые фразеологизмы. Разделимый фразеологизм – это фразеологизм, внутрь которого могут вставляться дополнительные слова. К примеру, внутрь словосочетания *под эгидой*, могут вставляться слова: *его*, *своей*, *одной*, *какой* и т.д.

Чтобы понять, надо ли учитывать, когда фразеологизмы используются в прямом значении, а когда в переносном, был проведен эксперимент на основе национального корпуса русского языка [8]. В результате оказалось, что для некоторых фразеологизмов, если внутрь словосочетания вставляются дополнительные слова, то они в большинстве случаев перестают быть фразеологизмом. Следовательно, их надо рассматривать как неразделимые. К таким относится словосочетания *белая ворона, адвокат дьявола*.

Разделимые фразеологизмы бывают двух типов. Первые – используются в значении фразеологизма, даже если внутрь попадают посторонние слова. К

примеру, *прыгнуть выше (своей, собственной) головы*. Однако вторые могут использоваться и в переносном значении, и в прямом. К таким фразеологизмам относятся *мутить воду, умывать руки, переходить дорогу* и так далее. Они все относятся к глагольной группе, поэтому в данной работе такие фразеологизмы рассматриваться не будут. Однако эта проблема требует отдельного решения.

Изменяемость словосочетаний

Фразеологизмы бывают изменяемые и неизменяемые. К неизменяемым фразеологизмам относятся, к примеру, *держи карман шире, по наклонной плоскости, дело в шляпе, не по зубам* и так далее. К изменяемым относятся фразеологизмы *согнуть в три погибели, волк в овечьей шкуре, блудный сын* и т.п.

Изменяемые фразеологизмы, в свою очередь, делятся на шесть групп в зависимости от типов изменяемых слов:

- существительное: *адвокат дьявола*;
- глагол: *бросить взгляд, бить баклуши*;
- прилагательное: *седой как лунь*;
- прилагательные + существительное: *белая ворона, абсолютный нуль*;
- числительное + существительное: *три грации*;
- существительное + существительное: *змий искуситель*.

Кроме фразеологизмов дополнительные смысловые нагрузки несут также словосочетания, являющиеся научными терминами (*математический анализ*) и именами собственными (*Баба Яга*), географические названия (*Великие Луки*). Такие словосочетания попадают в группы (прилагательные + существительное) и (существительное + существительное). Структура словаря устойчивых словосочетаний

Словарь составлен на основе словаря фразеологизмов [8]. Для удобства он поделен на две части – изменяемые и неизменяемые словосочетания.

Для каждого устойчивого словосочетания в словаре хранится следующая информация:

- Само словосочетание – с маркерами изменяемости слов и ограничениями на отдельные слова. Изменяемые слова представлены в начальной форме.
- Ключевое слово – главное слово в словосочетании (в предложной группе ключевым словом является существительное).
- Внутренняя группа – независимая синтаксическая роль словосочетания.
- Внешняя группа – синтаксическая роль словосочетания в рамках предложения.
- Разделимость – маркер разделимости фразеологизма.

Ограничения на грамматические характеристики фрагментов словосочетаний

В словаре после некоторых слов в скобках перечисляются их допустимые характеристики: род, падеж, вид и т.п. Например, *дело (nominative) в шляпе (prepositional,female).* Эти характеристики используются на этапе предсинтаксического анализа при свертке словосочетания в целом для сокращения вариантов разбора.

Например, в фразеологизме *вдохнуть жизнь*, “жизнь” стоит в винительном падеже, но морфологический анализатор вернет два варианта падежа: И.п. и В.п., тогда И.п. можно отбросить; или, есть ряд фразеологизмов, которые будут фразеологизмами только при фиксированном значении некоторых характеристик изменяемых слов (*крылатые слова* – только во множественном числе, *белая ворона* – только в женском роде).

Описание синтаксических групп

С каждым словосочетанием связывается его синтаксическая группа. Выделено пять синтаксических групп фразеологизмов:

- NP – именная группа, главное слово – существительное;
- VP- глагольная группа, главное слово – глагол;
- AP – группа прилагательного, главное слово – прилагательное;
- AdvP – наречная группа, главное слово – наречие;

- PP – предложная группа, главное слово – предлог.

В словаре фразеологизмов для каждого фразеологизма указаны две синтаксические группы – внутренняя и внешняя. Это требуется для анализа, например, следующих словосочетаний: *во всю прыть, во всю мочь, без ведома, сломя голову* и т.д.

Внешняя группа используется при анализе всего предложения, содержащего фразеологизм, она синтаксическую роль словосочетания в рамках предложения. Например, во фразе *бежал во всю прыть, во всю прыть* – это обстоятельство действия, а не дополнение.

Внутренняя же группа используется для проверки синтаксической правильности самого словосочетания. В первую очередь это нужно для разделимых фразеологизмов, поскольку не всегда можно предсказать, какие языковые конструкции окажутся внутри словосочетания. При этом словосочетание анализируется с помощью расширенной сети переходов, рассмотренной ниже и описывающей соответствующую синтаксическую группу, а также с учетом ограничений на характеристики отдельных слов словосочетания. Это позволяет убрать, например, неподходящие варианты разборы слова «*собственной»*, вставленного внутрь фразеологизма *прыгнуть выше головы*, которое имеет четыре варианта морфологического анализа в зависимости от падежа: Р.п., Д.п., Т.п. и П.п.

Примеры словарных статей устойчивых словосочетаний

Таблица 1.

Статья из словаря неизменяемых словосочетаний

Устойчивое словосочетание	**Ключевое слово**	**Внешняя группа**	**Внутренняя группа**	**Разделимость**
Во всю прыть (accusative)	Прыть	AdvP	PP	divisible

На примере данной статьи видно, что, во-первых, словосочетание будет играть роль наречия в предложении, но является именной группой, что может понадобиться для анализа вставленных внутрь фразеологизма слов. Во-вторых, морфологический анализатор построит два варианта разбора для слова "*прыть*" – в И.п. и в Р.п. А на основе словаря можно предсказать, какой падеж требуется в данном случае. В итоге это позволит значительно сократить количество вариантов разбора всего входного сообщения и убрать заведомо неверные.

Таблица 2.

Статья из словаря изменяемых словосочетаний

Устойчивое словосочетание	**Ключевое слово**	**Внешняя группа**	**Внутренняя группа**	**Разделимость**
Крылатый* (plural) слово* (plural)	Слово	NP	NP	undivisible

На примере этой статьи видно другое применение ограничений, которое может быть использовано еще на этапе поиска устойчивых словосочетаний: в единственном числе это не фразеологизм.

Предсинтаксический анализ устойчивых словосочетаний

На начальном этапе морфологический анализатор [???] получает всевозможные варианты разбора словоформ. На этапе предсинтаксического анализа выявляются устойчивые словосочетания и преобразуются в особую структуру, чтобы отличать эти словосочетания от других слов предложения и упростить дальнейшую работу синтаксического анализатора. При этом в описанную выше структуру добавляются следующие поля:

- RMUResult *nextWord – указывает на следующее слово словосочетания;
- bool mainWord – обозначает главное слово в словосочетании;
- int Phrase_cl – показывает внешнюю синтаксическую группу;
- int innerPhrase_cl – показывает внутреннюю синтаксическую группу.

В результате устойчивое словосочетание исходного сообщения заменяется на одно синтезированное слово, которое ссылается на первое слово словосочетания. Первое слово ссылается на второе, второе на третье и так далее. В эту цепочку включаются и вклинившиеся в устойчивое словосочетание слова, если оно разделимое. Главное слово словосочетания маркируется. Синтаксический класс синтезированного слова определяется по внешней группе соответствующей структуры.

Например, для фразы – *речь преимущественно идет о газетной утке*, которая содержит два фразеологизма: *речь идет* и *газетная утка* в результате работы предсинтаксического анализа будет получена следующая структура:

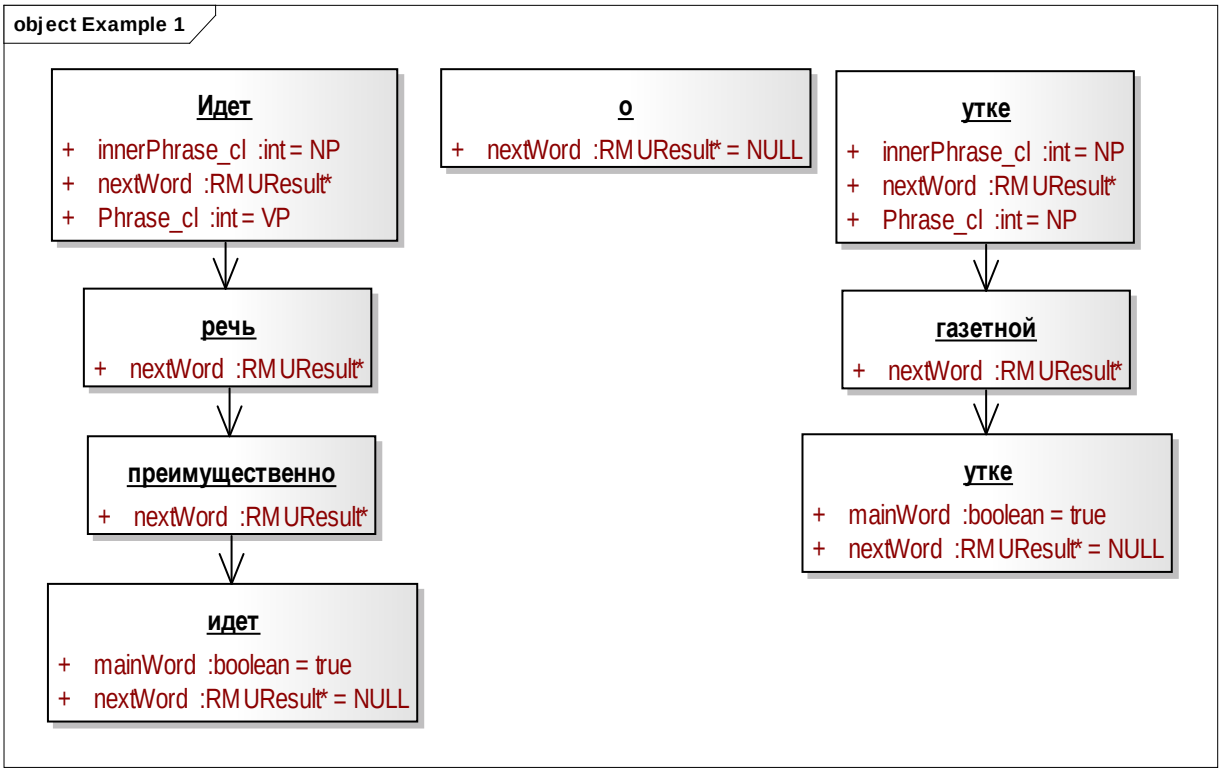

Рис. 1. Результат работы предсинтаксического анализа

Вместо фразеологизма *речь идет* синтезировано слово, хранящее информацию о главном слове фразы – *идет*. Далее оно ссылается на первое слово фразеологизма – *речь*, которое в свою очередь ссылается на следующее слово *преимущественно*, а оно - на слово *идет*. В словаре у слова *речь* стоит

ограничение – И.п., поэтому в итоге вместо двух вариантов разбора будет только один.

Аналогичная структура у фразеологизма *газетная утка*. Но у слова *газетной* четыре варианта разбора, а у слова *утке* – два. Однако при построении итоговой структуры существительное и прилагательное согласуются между собой, и в итоге получается только два варианта словосочетания – в Д.п. и в П.п.

А если подать на вход фразу – *чувствуешь себя не в своей тарелке*, содержащую неизменяемый фразеологизм *не в своей тарелке*, то получим такую структуру:

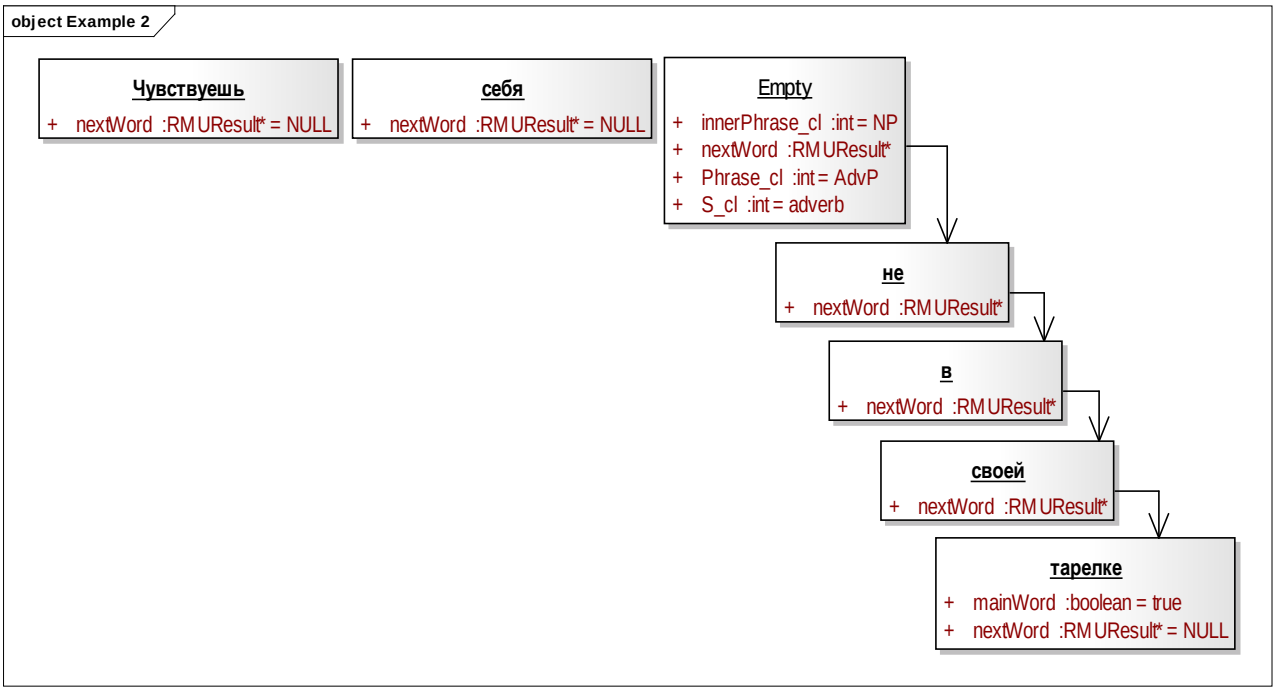

Рис. 2. Результат работы предсинтаксического анализа

В этом сообщении аналогично первому примеру происходит свертка фразеологизма, но так как внутренняя группа и внешняя отличаются, то синтезируется пустое первое слово, у которого обозначена только часть речи – наречие (AdvP).

Синтаксический анализ фраз русского языка с учетом устойчивых словосочетаний

Для синтаксического анализа фраз русского языка реализован модифицированный алгоритм Кока-Янгера-Касами на основе расширенной сети переходов [9, 10], позволяющий по заданной входной цепочке слов как выделить все допустимые именные группы, так и провести синтаксический анализ исходной фразы в целом.

Примеры работы анализатора

1) На вход анализатору подано сообщение: *часто применяемый в клинике препарат*. В результате анализа сетью для простого предложения получилось четыре варианта разбора вида, представленного на рис.3.

При этом ключевая именная группа состоит из главного слова – *препарат* и зависимой причастной группы – *часто применяемый в клинике*.

Четыре варианта получаются из-за того, что слова *препарат* и *применяемый* могут иметь И.п. или В.п.

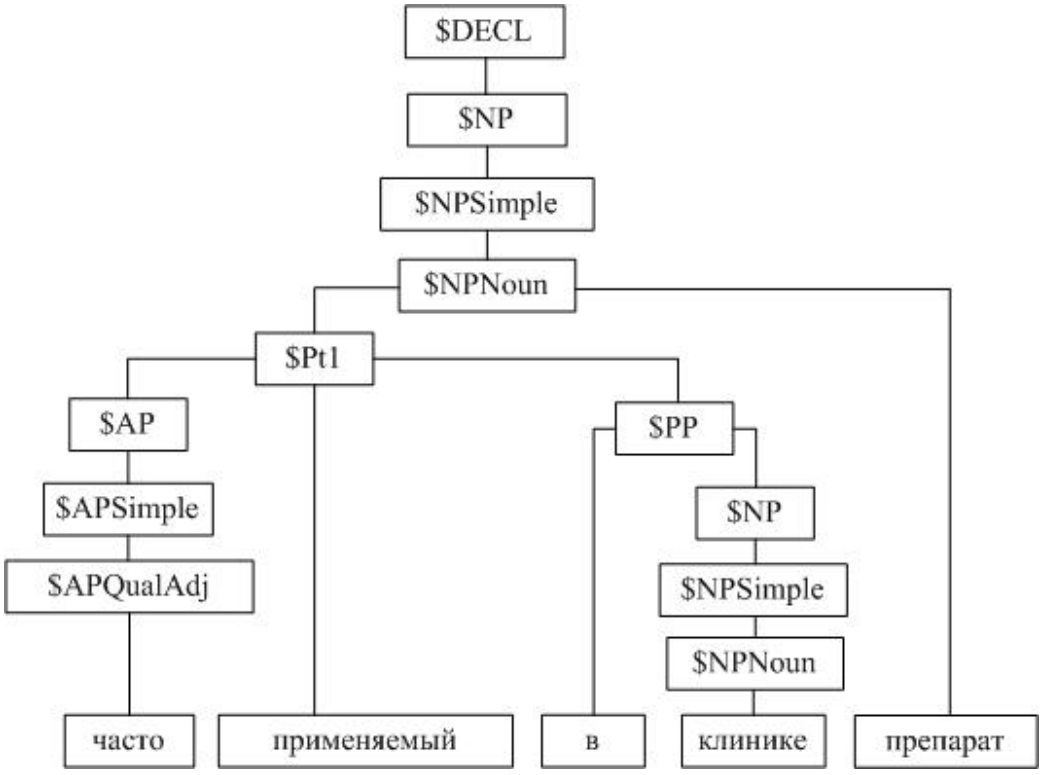

Рис. 3. Результат синтаксического анализа

2) На вход анализатору подано сообщение: *змей искуситель окончил московский государственный университет*. В результате при запуске без выделения устойчивых словосочетаний получилось двадцать шесть различных вариантов разбора. Если же запустить синтаксический анализатор после этапа предсинтаксического анализа, то получится всего три варианта, поскольку *Московский Государственный Университет* и *змей искуситель* являются устойчивыми словосочетаниями, что существенно улучшает качество разбора.

На рисунке 4 изображена схема разбора данной фразы. Ключевая именная группа выражена синтезированной на предыдущем этапе предсинтаксического анализа структурой *московский государственный университет*. Также есть глагольная группа, выраженная глаголом *окончил* и дополнением из второго фразеологизма *змей искуситель*.

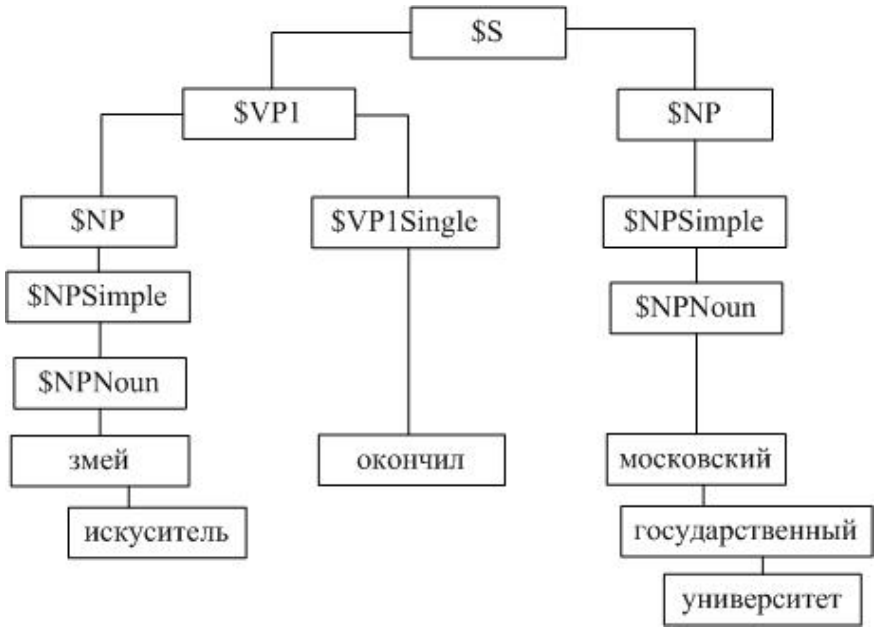

Рис. 4.Результат синтаксического анализа. Главное слово - университет

На рисунке 5 изображена схема разбора того же входного сообщения, но в данном случае ключевая именная группа – синтезированный фразеологизм *змей искуситель*. Второе устойчивое словосочетание – *московский государственный университет* входит в состав глагольной группы в качестве дополнения. К такой

схеме относятся два варианта разбора, так как *московский государственный университет* может иметь И.п. или В.п.

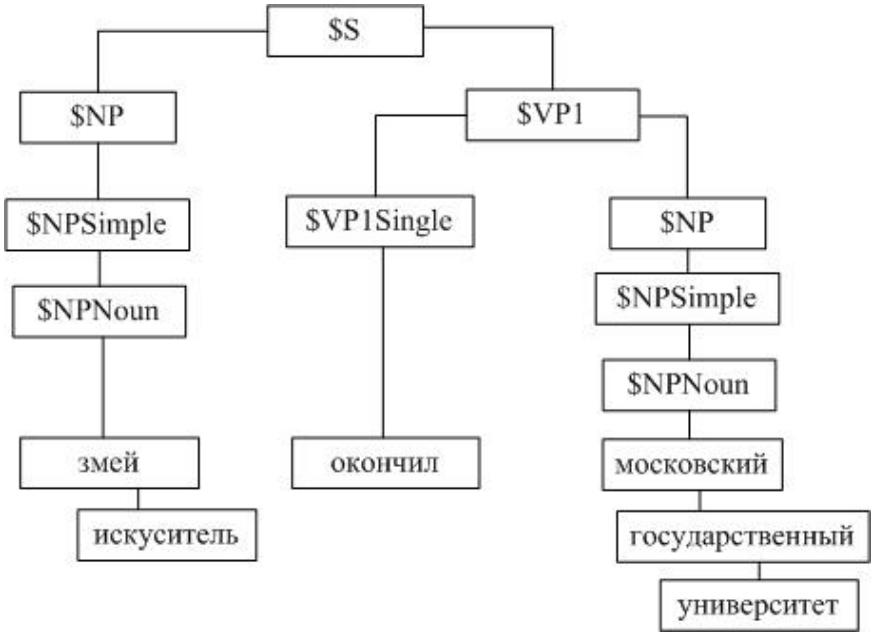

Рис. 5. Результат синтаксического анализа. Главное слово - змей

Результаты. Обсуждение и анализ.

Для оценки качества работы синтаксического анализатора, учитывающего устойчивые словосочетания, были выбраны 800 сообщений (на основе Национального корпуса русского языка [8]), содержащих неизменяемые и изменяемые фразеологизмы. В результате получили, что за счет выделения устойчивых словосочетаний количество разборов всегда не увеличивает количество разборов без выделения устойчивых словосочетаний.

Оценка качества работы синтаксического анализатора

На рисунках 6 и 7 приведены графики количества различных потенциально правильных вариантов разбора с предсинтаксическим анализом устойчивых словосочетаний и без него.

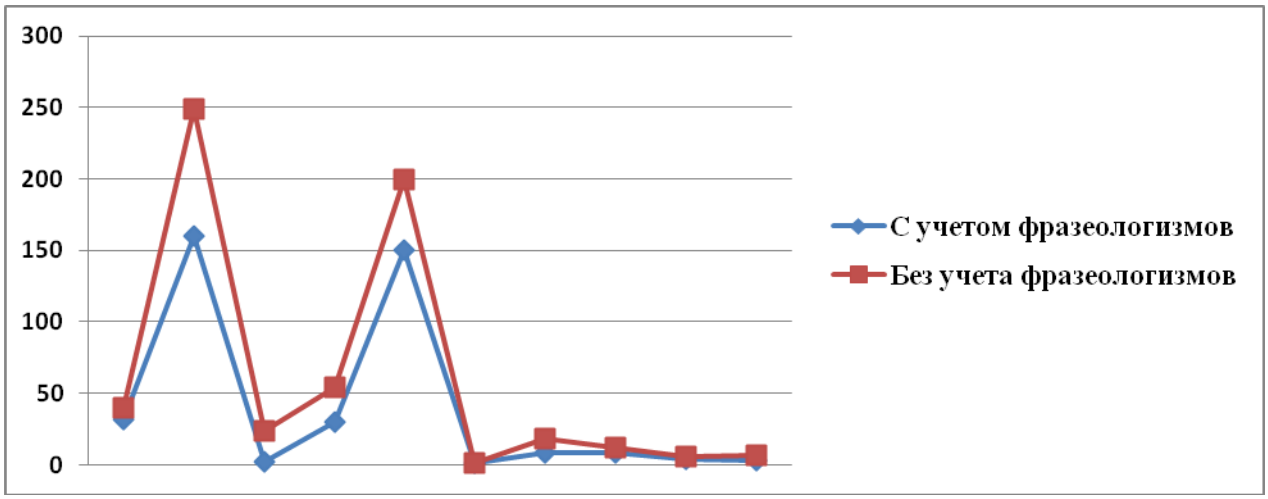

Рис. 6. Диаграмма количества вариантов разбора для неизменяемых фразеологизмов

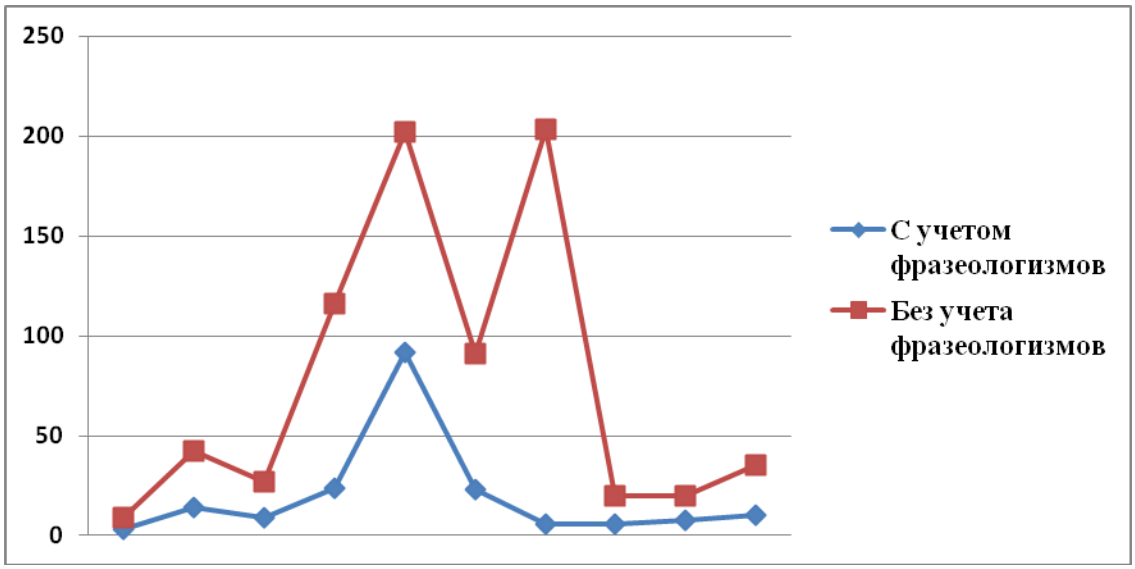

Рис. 7. Диаграмма количества вариантов разбора для изменяемых фразеологизмов

Заключение и выводы.

Были рассмотрены возможности повышения качества синтаксического анализа фраз русского языка за счет учета и предварительного анализа устойчивых словосочетаний.

Проанализированы свойства и структура различных устойчивых словосочетаний, разработан словарь устойчивых словосочетаний, учитывающий

роль словосочетаний в предложении и ограничения на характеристики отдельных его слов. Полученная структура устойчивых словосочетаний позволит далее на этапе семантического анализа учитывать специфический смысл фразеологизмов.

Реализован метод поиска устойчивых словосочетаний на этапе предсинтаксического анализа, позволяющий выделить такие словосочетания в отдельную подструктуру и анализировать их далее как одно слово и тем самым сократить количество неверных вариантов разбора.

Кроме того, разработана расширенная сеть переходов для грамматики русского языка, позволяющая, в частности, выделять ключевые именные группы из входного сообщения.

В результате предсинтаксический анализ позволил значительно сократить количество потенциально правильных вариантов разбора входных сообщений, содержащих устойчивые словосочетания.

В дальнейшем можно развить предложенный метод, сворачивая на этапе предсинтаксического анализа сложносоставные числительные, составные глаголы и тому подобное. Разработанная расширенная сеть переходов может быть использована и для построения более сложных систем, основанных, к примеру, на моделях управления.

Литература:

1. Волкова И.А. Введение в компьютерную лингвистику. Практические аспекты создания лингвистических процессоров: Учебное пособие для студентов факультета ВМиК МГУ. – Москва: Издательство МГУ, 2006. – с.44.

2. Serge Sharoff, Joakim Nivre. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge // Dialog 2011. Computational Linguistics and Intellectual Technologies. (International Conference. Moscow, RGGU Publishers, Issue 10(17). P. 591-604.

3. Leonid Iomdin, Vadim Petrochenkov, Victor Sizov, Leonid Tsinman. ETAP parser: state of the art // Dialog 2012. Computational Linguistics and Intellectual Technologies. (International Conference. Moscow, RGGU Publishers, Issue 11(18). P. 830-843.

4. Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova M. A., Selegey V. P., Zuev K. A. // Dialog 2012. Computational Linguistics and Intellectual Technologies. (International Conference. Moscow, RGGU Publishers, Issue 11(18). P. 810-822.

5. Антонова А. А., Мисюрев А. В. Об использовании синтаксического анализатора Cognitive Dwarf 2.0 // Труды ИСА РАН. Т. 38, 2008, с. 91-109.

6. Федосов И.В., Лапицкий А.Н. Фразеологический словарь русского языка. – Москва: Юнвес, 2003. – с.608

7. Поспелов Е.М. Географические названия мира. Топонимический словарь. – Москва: Русские словари, 1998. – с.504

8. Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л. и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка: 2003—2005. М.:Индрик, 2005, 193—214.

9. И.А.Волкова, И.Г.Головин. Об одном подходе к построению синтаксического модуля в системе распознавания устной речи. // Труды Международного семинара Диалог-97 по компьютерной лингвистике и ее приложениям. – Москва, 1997. с. 61-63.

10. Волкова И.А., Головин И.Г. Синтаксический анализ фраз естественного языка на основе сетевой грамматики. // Труды Международного семинара Диалог-98 по компьютерной лингвистике и ее приложениям. – Москва, 1998. с. 438–447.

Статья отправлена: 03.12.2015 г.