СЕМАНТИКО-СИНТАКСИЧЕСКИЙ АНАЛИЗАТОР SEMSIN
THE SEMANTIC-AND-SYNTACTIC PARSER SEMSIN

Каневский Е.А. (kanev@emi.nw.ru), Санкт-Петербургский экономико-математический институт РАН,

Боярский К.К. (boyarin9@yandex.ru), Национальный исследовательский университет информационных технологий, механики и оптики

Санкт-Петербург, Россия.

Аннотация. Описан принцип работы семантико-синтаксического анализатора SemSin, строящего дерево зависимостей для предложений русского языка с помощью набора продукционных правил. Приведена последовательность применения блоков правил, рассмотрены примеры их работы. Особенностью правил является принятие решений об установлении синтаксических связей с одновременным снятием морфологической омонимии. Показано, что активное использование синтаксической и семантической словарной информации позволяет значительно уменьшить неоднозначность разбора.

Ключевые слова. Парсер, лексема, дерево зависимостей, синтаксический анализ, продукционные правила

Kanevsky E.A. (kanev@emi.nw.ru), St. Petersburg Institute for Economics and Mathematics, RAS

Boiarsky K.K. (boyarin9@yandex.ru), National Research University of Informational Technologies, Mechanics and Optics

Abstract. The principle of work of the semantic-and-syntactic parser SemSin is described. This system fulfils morphological and syntactical analysis of Russian sentences and builds the dependence tree for it. For the description of morphologic, syntactic and semantic characteristics of words the dictionary and the classifier are used. Analyzer functioning is carried out by means of a set production rules. The general characteristic of principles of construction of these rules is given. Each rule is divided on conditional and executive parts. The operators used for construction of a conditional part of rules, provide the analysis of all morphological and graphematical parameters of separate tokens, and also their relative positioning. Besides, actants and semantic classes of the lexemes are analyzed. The commands used for construction of an executive part of rules, provide links between separate tokens and ambiguity reduction. The sequence of application of rule's blocks is given, examples of their

work are considered. Feature of rules is the decision making about an establishment of syntactic links to simultaneous removal of a morphological ambiguity. It is shown that the active use of the syntactic and semantic dictionary information allows reducing ambiguity of analysis considerably.

Key words. Parser, token, lexeme, depending tree, syntactic analysis, production rules

Введение

Компьютерная морфология необходима в прикладных системах, ведущих поиск и анализ информации на естественном языке. Вопросам компьютерной морфологии посвящено множество работ, однако как показал проведенный в 2010 г. форум «Оценка методов АОТ», эта проблема до сих пор не решена окончательно [Ляшевская 2010]. Оценивая результаты работы нашего морфолого-лексического анализатора TextAn [Каневский, Боярский 2010], производящего морфологический разбор с последующим снятием омонимии и уменьшающего морфологическую омонимию до уровня порядка 10%, мы пришли к выводу, что для ее существенного снижения необходимо производить синтаксический разбор (возможно с использованием элементов семантики). Это и послужило причиной, побудившей авторов к разработке семантико-синтаксического анализатора SemSin, который должен не только снимать частеречную и морфологическую омонимию, но и строить синтаксическое дерево зависимостей, а возможно, даже частично снимать лексическую неоднозначность.

Анализатор участвовал в форуме синтаксических парсеров, проводившемся в ноябре 2011 г. Было разобрано около 65 тыс. предложений. Никакой специальной подготовки программы к конкурсу не производилось за исключением согласования входного и выходного форматов. После получения исходных текстов было произведено пополнение словаря наиболее часто встречающимися словами.

В качестве исходных лексических материалов используются словарь и классификатор В. А. Тузова [Тузов 2004]. Словарь Тузова основан на морфологическом словаре А. А. Зализняка [Зализняк 1980], при определении его семантики широко использовался словарь С. А. Кузнецова [Кузнецов 1998]. По существу, к настоящему времени в исходный словарь добавлено более 40 тыс. слов общей лексики и более 20 тыс. названий (стран, городов, учреждений и фирм, рек, морей и т. п.) и собственных имен. Все 177 тыс. лексем распределены по 1660 классам. Неоднозначность слов такова – около 14% слов имеет две и более лексемы, которые в большинстве случаев относятся к разным классам (классический пример: три лексемы для слова *коса*).

В процессе разработки исходный словарь был упрощен – на его основе создана морфологическая база данных. В ней каждая лексема содержит морфологические характеристики, а также номер своего класса и актанты вызываемых ею лексем в виде падежей или предлогов с соответствующими падежами – например, вВин, вПред и т. д. Часто перед таким актантом указаны допустимые классы слов, могущих их замещать. В случае если несколько лексем одного и того же слова имеют одинаковое морфологическое описание, они объединяются в одну лексему. В этом случае неоднозначность составит около 3%, что приблизительно соответствует неоднозначности в словаре Зализняка. Морфологический анализатор, используя эту базу данных, выдает результат разбора очередной словоформы в виде леммы (слова в нормальном виде) с морфологическими характеристиками и класса (или набора классов) с указанием соответствующих актантов.

Результаты разбора каждого слова хранятся в отдельном элементе, в котором может размещаться до 7-ми лексем с разными морфологическими характеристиками. Интересно, что из всех встретившихся нам на сегодня слов наибольшим количеством «разнотипных» лексем, по-видимому, обладает словоформа *боров*. Ниже для каждой ее лексемы указано название класса по классификатору, в фигурных скобках приведены параметры слова по словарю А. А. Зализняка; при этом первой строке соответствуют две лексемы, но поскольку они обладают одинаковыми морфологическими параметрами, то и объединяются в один элемент:

БОР м1 Мн. Род. Хим. эл. + Инстр. Мед. {БОР м1а (инструмент; хим. элемент)}

БОР м1В Мн. Род. Ландшафт Лес {БОР м1с, $П_2$(в) (лес)}

БОР м1о Мн. Род. Вин. Личность ФИО Фамилия { }

БОРЫ м1+ Мн. Род. Деньги Взыск Налог {БОРЫ мн.<м1b>}

БОРОВ м1о Ед. Им. Млекопит. Домашние Свиньи {БОРОВ мо1е (кабан)}

БОРОВ м1|1 Ед. Им. Вин. Огонь Спец_место {БОРОВ м1с|1 (часть дымохода)}

Следует заметить, что SemSin, как правило, анализирует только те слова, которые имеются в словаре, и рассчитан на разбор грамотно написанных текстов.

В анализаторе достаточно эффективно решается проблема словосочетаний. Полученная из исходного словаря специальная база фразеологизмов обеспечивает разбор трех типов словосочетаний: неизменяемых (*несмотря ни на что, вдалеке от*), с изменяемым первым словом (*гвоздь программы*) и полностью изменяемых (*белая ворона*). В настоящее время база содержит более 4100 фразеологизмов и играет важную роль в снятии неоднозначности, особенно для составных предлогов, союзов и наречий. Используется также отдельная база предлогов, хранящая классы существительных, с которыми они взаимодействуют, и названия связей с хозяевами предложных групп.

Особенности составления правил

На вход анализатора подается текст на русском языке по абзацам. После выделения токенов текст подвергается морфологическому анализу. При этом каждому токену приписывается не только морфологическая, но и синтаксическая (актанты), и семантическая информация. Затем цепочка токенов обрабатывается с помощью системы продукционных правил с целью снижения неоднозначности и перехода от линейной схемы предложения к синтаксическому дереву. Система делает попытку применения каждого из правил последовательно ко всем токенам и при выполнении указанных в правиле условий совершает определенные действия.

Принятый нами за основу язык правил уточнялся в процессе разработки системы. Общее количество типов используемых операторов и команд более 100. Язык правил, по нашему мнению, достаточно адекватен как решаемой задаче, так и условию простоты отладки [Боярский, Каневский 2011]. Сейчас составлено более 210 правил.

Каждое правило начинается с **условной части**, которая строится по обычной для языков программирования схеме:

If…Then…ElseIf…Then…Else…EndIf.

Внутри каждого блока могут использоваться операторы конъюнкции & и дизъюнкции OR. Если какое-то условие в блоке не выполнено, то остальные не проверяются. Разрешаются вложенные условия, но не более двух уровней.

Первая группа операторов определяет позицию токена в абзаце или предложении. Вторая группа операторов проверяет тип токена: знак пунктуации, наличие цифры, регистр. Следующая группа анализирует морфологические характеристики слова. Поскольку одному токену могут соответствовать несколько разных лексем, а каждой лексеме – различные грамматические параметры (падеж, число, время и др.), эти операторы, возвращают значение TRUE, если условие выполняется хотя бы для одной лексемы.

Четвертая группа операторов проверяет полное или частичное согласование слов по роду, числу и падежу. В частности, проверяется согласование прилагательных (причастий и порядковых числительных) с существительными, подлежащего со сказуемым. Отдельная группа операторов проверяет согласование актантов данной лексемы с подключаемыми к ней словами. Наконец последняя группа анализирует фрагменты уже созданного синтаксического дерева, в частности, наличие и имя входных связей именных групп или отдельных токенов. Всего на данный момент в условной части правил использовано более 50 типов операторов проверки условия.

Если все условия проверявшейся ветви правила удовлетворены, выполняются команды **исполнительной части**. Это, во-первых, команды, уменьшающие неоднозначность разбора путем удаления лексем с заданными характеристиками. Ряд других команд согласуют слова по роду, числу и падежу, оставляя только согласованные варианты токенов. Имеются команды, выделяющие именные группы и устанавливающие связи между словами.

Несколько команд служат для сегментации предложения. Обычно центром сегмента является сказуемое (глагол, краткое прилагательное или причастие, предикат) в главных или придаточных предложениях, деепричастие или причастие в соответствующих оборотах. Имеются команды для выделения сегмента, для установления его центра в определенную позицию, для объединения сегментов. Работа с сегментами производится по правилам, близким к описанным Т.Ю. Кобзаревой [Кобзарева 2004]. Всего в исполнительной части насчитывается более 50 команд.

Специфика задачи породила ряд особенностей составления этих правил. В основу разработки было положено несколько гипотез.

1. Задачи снятия морфологической омонимии и построения синтаксического дерева настолько взаимно переплетены, что решать их нужно одновременно.

2. Система правил не может быть представлена в виде простого набора конъюнкций условий, так как результат зависит от порядка применения правил.

3. Зона действия многих правил не локальна, поэтому целесообразно различать их по способу применения – master (правило, последовательно анализирующие все токены заданного сегмента) и slave (рекуррентное правило, анализирующее заданную окрестность исходного токена). Master-правила исполняются в порядке их расположения, slave-правило запускается только при выполнении условной части master-правила, его вызывающего. Для удобства составления и контроля последовательности их выполнения правила делятся на группы, каждая их которых представляет собой обычный текстовый файл.

Принятая схема применения правил обеспечивает независимость скорости работы анализатора от длины предложения. Время анализа линейно связано с общим числом поданных на разбор слов.

Далее мы обсудим принятый в системе SemSin порядок выполнения правил и рассмотрим примеры их функционирования. Следует отметить, что оптимальная последовательность применения правил не выводится нами из каких-либо теоретических рассуждений, а уточняется в процессе экспериментов. Часть анализируемых предложений и их фрагментов взяты из корпуса текстов конкурса синтаксических парсеров 2011 г.

Еще раз подчеркнем, что основным результатом работы анализатора SemSin мы считаем правильно построенное дерево зависимостей с однозначно определенными морфологическими параметрами узлов. На основе этих данных возможна дальнейшая обработка, проводимая с различными целями: семантическая разметка, построение онтологий и сценариев и т. д. Поскольку в этом смысле результаты SemSin носят промежуточный характер, мы не стремились к унификации названий связей с какой-либо из известных систем анализа, тем более что стандарта де-факто в этой области не существует. Большинство связей в нашей системе именуются либо по падежам подключаемого слова (*прочитал* →[Вин] *газету*), либо по сочетанию предлог-падеж (*живу* →[наВин] *на*

средства), либо в соответствии с теми вопросами, которые к ним можно поставить (*живу* →[Где] *на море*).

1. Выделение токенов

После считывания очередного абзаца текста морфологический анализатор разбирает токены и приписывает им соответствующие морфологические характеристики. На этом этапе выделение токенов производится по пробелам и знакам препинания. Однако в ряде случаев лексические единицы в предложении имеют более сложную структуру. Первая группа правил и производит синтез таких сложных токенов. это могут быть, например, сложные единицы измерения (*кв.м, км/час*), интернет-адреса (*http://yandex.ru*), записанные цифрами порядковые числительные (*1917-й*).

Одновременно по словарю фразеологизмов объединяются в один токен неизменяемые слова, образующие сложные наречия, предлоги и т. д. (*несмотря ни на что, вдалеке от*). Слова, входящие в состав изменяемых фразеологизмов (*гвоздь программы, белая ворона*) в единый токен не объединяются, но сразу соединяются в лексическую группу подходящей связкой. При этом группе приписывается соответствующий семантический класс. Так *белая ворона* получит класс людей, а *атомное сердце* не будет иметь отношения к органам человеческого организма.

Затем производится выявление инициалов и аббревиатур и подключение их к словам-хозяевам (рис. 1).

рис. 1.

В результате становится ясно, что некоторые точки не являются границами предложений, несмотря на то, что после них стоит прописная буква. Только после этого выполняется разбивка абзаца на предложения, и весь дальнейший анализ ведется уже только в пределах одного предложения [Боярский, Каневский 2010].

2. Обработка отдельных слов

Некоторые часто встречающиеся слова русского языка обладают высокой степенью морфологической омонимии. В ряде случаев анализ ближайшего окружения позволяет снизить эту степень и тем самым упростить дальнейший разбор. К таким словам относятся, например, *это* (леммы ЭТО или ЭТОТ), *тем* (ТОТ–ТЕМА–ТЕМ), *потом* (ПОТ–ПОТОМ), *перед* (сущ.–предлог), *по* (предлог – название реки – аббревиатура [программное обеспечение]) и т. д. Так при разборе предложения

Перед Новым Годом по решению правления компания закупила двадцать две лицензии на ПО

Правила этой группы должны сработать три раза для разрешения омонимии:

Перед → предлог, так как следом стоит слово в Тв. падеже;

по → предлог, так как правее расположено слово, начинающееся со строчной буквы;

ПО → аббревиатура, так как следующее слово написано прописными буквами.

3. Выделение групп с фамилиями, названиями, числами

В предложении

Капитан Джон Сильвер приехал на остров Ямайка 12 мая 1730 года

образуются три лексические группы:

(*Капитан Джон Сильвер*), поскольку слово из класса профессий соседствует и согласуется с именами собственными;

(*остров Ямайка*), поскольку слово из класса географических объектов соседствует с именем собственным;

и группа даты: (*12 мая 1730 года*).

Рассмотрим более подробно механизм выделения первой группы. Обнаружив в тексте слово, написанное с прописной буквы, правило анализа прописных букв устанавливает, что соответствующая лексема имеется в словаре и принадлежит классу имен. Затем запускается подчиненное правило, которое находит справа фамилию (а по возможности и отчество). Следующее подчиненное правило пытается найти слева слово из класса профессий. Это правило рекуррентно вызывает само себя с последовательным сдвигом влево до тех пор, пока не найдет подходящее слово. Поиск прерывается, если очередной токен является знаком препинания или глаголом. Именно многообразие условий прерывания поиска привело нас к идее подчиненных правил.

Из приведенных выше примеров видна важность словарной семантической информации, без которой выделение лексических групп было бы гораздо более сложным.

4. ***Подключение прилагательных и причастий, снятие неоднозначностей прилагательное–существительное***

На этом этапе система, обнаружив прилагательное или причастие, начинает нелокальный поиск согласованного существительного вправо и влево. При этом учитывается словарная информация о том, какие классы существительных подходят. Например, встретив сочетание *моющие средства*, определяется, что лемма существительного – СРЕДСТВО, в то время как сочетание *денежные средства* дает другую лемму – СРЕДСТВА. Таким образом, семантическая информация способствует снижению морфологической омонимии.

В этой же группе правил, по возможности, снимается омонимия прилагательное–существительное: *школьная столовая* vs *столовая ложка*.

5. Подключение предлогов

Подключение предлогов происходит в два этапа. Сначала образуются предложные группы. Предложные группы могут быть вложенными как, например, в предложении

Она долго еще принаряживалась и жеманилась перед небольшим ***в оловянных рамках*** *зеркалом*

Во избежание конфликтов просмотр предлогов производится справа налево. В зависимости от того, какой предлог с каким именно существительным и в каком падеже связывается, определяются возможные типы связей. В данном случае это связи *вПред* (по предложному падежу), *Какой, Где* для предлога *в* и связи *Перед, Где* для предлога *перед*.

Поиск хозяина предложной группы выполняется после выделения сказуемых и сегментации предложения (см. ниже). Это необходимо для того, чтобы сузить зону поиска хозяина. Делается попытка подключить предложную группу к предикативной вершине. В рассматриваемом случае синтаксическая словарная информация ничем не может помочь: подходящих актантов у обоих глаголов нет. Однако остается возможность подключить предложную группу по свободной связи, т. е. по такой, которая допустима для любого глагола. К свободным связям в нашей системе относятся, например *Где, Как, Когда* и некоторые другие. Таким образом, происходит первое подключение (рис. 2):

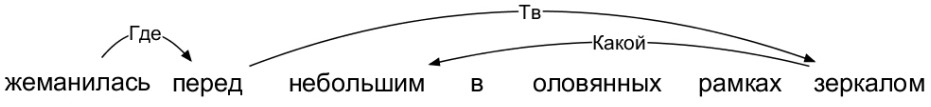

рис.2.

Группа предлога *в* уже не может подключаться к глаголу, так как это привело бы к нарушению проективности связей [Тестелец 2001]. Подходящих аргументов у соседних слов нет, поэтому эта предложная группа подключается по свободной связи *Какой* к ближайшему существительному *зеркалом* (рис. 3).

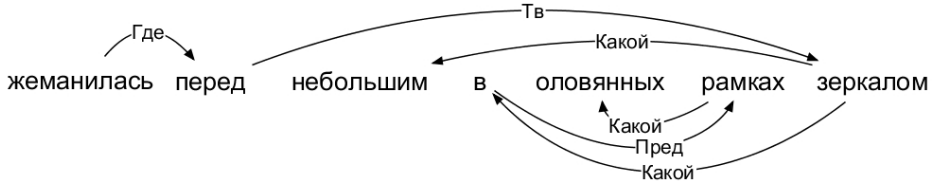

рис.3.

6. Сегментация

Работа с сегментами (обособленными оборотами, придаточными предложениями и т. д.) производится в основном по алгоритмам, предложенным Т.Ю. Кобзаревой [Кобзарева 2004]. Предложение разбивается по знакам препинания на отрезки, для каждого отрезка делается попытка найти предикативную вершину. В случае если такой вершиной оказывается неомонимичный глагол, у всех других слов отрезка удаляются предикатные омоформы. Затем производится обратное действие: выявление запятых, не являющихся границами сегментов (прежде всего между однородными членами) и слияние отрезков в сегменты.

7. Поиск составных сказуемых и подлежащих

Составные сказуемые, особенно имеющие в качестве присвязочной части инфинитив, обычно достаточно легко компонуются с помощью актантов глаголов, указанных в словаре. Рассмотрим предложение

Тем более их не могли бы выдержать члены экипажей пилотируемых космических аппаратов

Глагол МОЧЬ имеет актант Инфин и образует составное сказуемое *могли* → *выдержать*. Однако затем нужно определить, что подлежащее *члены* согласуется с основным глаголом, а местоимение *их* должно подключиться к переходному глаголу *выдержать*.

В ряде случаев оставшаяся к данному моменту морфологическая неоднозначность серьезно затрудняет поиск подлежащего. Так в предложении

Ветви ивы нависли над самой водой

у слова *ветви* возможны Род., Дат. и Пред. пад. Ед. ч. и Им. и Вин. пад. Мн. ч., а у слова *ивы* – Род. пад. Ед. ч. и Им. и Вин. пад. Мн. ч. Таким образом, оба эти слова являются кандидатами на роль подлежащего, но слово *ивы* как будто предпочтительнее – оно ближе к глаголу. Сделать правильный выбор здесь снова помогает словарная информация: слово *ветвь* способно присоединять существительное в Род. пад., принадлежащее к семантическому классу растений. Таким образом, устанавливается правильный порядок связей (рис. 4):

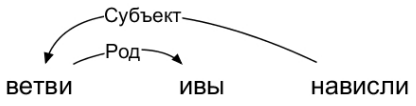

рис. 4.

Заодно устраняется морфологическая неоднозначность и снимается лексическая омонимия: в данном контексте слово *ветвь* обозначает часть растения, а не что-то другое.

8. Окончательная обработка дерева связей

На заключительном этапе производится подключение предложных групп, существительных, местоимений и числительных к хозяевам. Устанавливаются межсегментные связи (деепричастные, вводные, сложносочиненные, сложноподчиненные). Производится повторное согласование в группах существительное – прилагательное. Необходимость этого вызвана тем, что в процессе сборки синтаксического дерева морфологическая неоднозначность существительного могла уменьшиться.

После окончания работы правил результат разбора представляется в виде дерева на экране и выводится в виде XML-файла.

ЛИТЕРАТУРА

1. Боярский К.К., Каневский Е.А. Разбиение текста на предложения // Дискуссия теоретиков и практиков. Научно-практический журнал. 2010. №1 (3). С. 135–137.

2. Боярский К.К., Каневский Е.А. Язык правил для построения синтаксического дерева // Интернет и современное общество: Материалы XIV Всероссийской объединенной конференции «Интернет и современное общество». – СПб.: ООО «МультиПроджектСистемСервис», 2011. С. 233–237.

3. Зализняк А.А. Грамматический словарь русского языка. М: Русский язык, 1980.

4. Каневский Е.А., Боярский К.К. Морфолого-лексический анализатор и классификация текста // Прикладная лингвистика в науке и образовании. Материалы V международной научно-практической конференции 25–26 марта 2010. – СПб.: «Лема», 2010, с. 157–163.

5. Кобзарева Т.Ю. Принципы сегментационного анализа русского предложения // Московский лингвистический журнал. М.: РГГУ, 2004. Т.8, №1, с. 31–80.

6. Кузнецов С.А. Большой толковый словарь русского языка. – СПб.: Норинт, 1998.

7. Ляшевская О.Н. и др. Оценка методов автоматического анализа текста: морфологические парсеры русского языка // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 9 (16). М.: Изд-во РГГУ, 2010. С. 318–326.

8. Тестелец Я. Г. Введение в общий синтаксис. М.: РГГУ, 2001.

9. Тузов В.А. Компьютерная семантика русского языка. СПб.: Изд-во С.-Петерб. ун-та, 2004.

1. Boiarsky K.K., Kanevsky E.A. (2010), Splitting of the text into sentences [Razbienie teksta na predlozhtniia], Diskussiia teoretikov I praktikov. Nauchno-prakticheskii zhurnal [Discussion of theorists and experts. Scientifically-practical magazine], no. 1 (3). pp. 135-137.

2. Boiarsky K.K., Kanevsky E.A. Language of rules for construction of a syntactic tree [Iazyk pravil dlia postroeniia sintaksicheskogo dereva]. Internet I sovremennoe obshchestvo: Materialy XIV Vserossiiskoi ob"edinennoi konferentsii “Internet I sovremennoe obshchestvo” [*The Internet and a modern society: Materials of XIV All-Russia incorporated conference “The Internet and a modern society”*]. St.-Petersburg, 2011, pp. 233-237.

3. Kanevsky E.A., Boiarsky K.K. The morfologo-lexical analyzer and text classification [Morfologo-leksicheskii analizator I klassifikatsiia teksta]. Prikladnaia lingvistika v nayke I obrazovanii. Materialy V mezhdunarodnoi naychno-prakticheskoi konferentsii 25-26 marta 2010 [*Applied linguistics in science and education. Materials of V international scientifically-practical conference 25-26 march of 2010*]. St.-Petersburg, 2010, pp. 157-163.

4. Kobzareva T.Iu. (2004), Principles the segmented analysis of the Russian sentences [Printsipy segmentatsionnogo analiza russkogo predlozhtniia], Moskovskii lingvistichtskii zhurnal [The Moscow linguistic magazine], M.: РГГУ t. 8, no. 1, pp. 31-80.

5. Kuznetsov S.A. (1998), Bol'shoi tolkovyi slovar' russkogo iazyka [The big explanatory dictionary of Russian], Norint, St.-Petersburg.

6. Liachevskaia O.N. and other. Estimation of methods of the automatic analysis of the text: morphological parsers of Russian [Оценка методов автоматического анализа текста: морфологические парсеры русского языка]. *Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010"* [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010"]. ???, 2010, pp. 318-326.

7. Testelets Ia.G. (2001), Vvedenie v obshchii sintaksis [Introduction in the general syntax], RGGU, Moscow.

8. Tuzov V.A. (2004), Komp'iuternaia semantika russkogo iazyka [Computer semantics of Russian], St.-Petersburg University Publ., St.-Petersburg.

9. Zalizniak A.A. (1980), Grammaticheskii slovar' russkogo iazyka [The grammatical dictionary of Russian], Russian language, Moscow.