Синтактико-семантические методы снятия омонимии при распознавании речи

Лобанов Б.М., *доктор технических наук*

Житко В.А., *аспирант*

В статье описываются методы снятия омонимии при распознавании речи с использованием результатов синтаксического и семантического анализа текста. В качестве инструментального средства анализа текстов русского языка взят программный комплекс автоматической обработки текстов — АОТ.

• *распознавание речи* • *синтаксический анализ* • *семантический анализ* • *омонимия* • *омографы* • *омофоны*

The method for disambiguation in speech recognition using the results of the semantic analysis of the text is described. The software system of automatic natural language text processing is taken as a tool of analysis of Russian texts.

• *speech recognition* • *syntactic analysis* • *semantic analysis* • *homonymy* • *homographs* • *homophone*

Введение

При распознавании и понимании речи возникает ряд проблем, связанных с неоднозначностью естественного языка. Один из источников неоднозначности — омонимия — явление, при котором некоторые слова естественного языка могут иметь одинаковое написание или произношение.

Если слова имеют одинаковое написание, но различное произношение, они называются **омографы**, например «*за̀мок*» — «*замо̀к*», «*ду̀хи*» — «*духѝ*», «*вѐсти — вестѝ*», «*бо̀чка — бочка̀*», «*бѐрег — берёг*», «*сте́кла*» — «*стекла̀*». Такие слова чаще всего имеют различие в ударении или при написании буквы «ё» без точек [1]. Задача снятия омографии особо остро стоит при синтезе речи по тексту, так как неправильная интерпретация омографов может значительно ухудшить, а в ряде случаев и исказить смысл синтезируемого текста [2].

Если слова, напротив, имеют одинаковое произношение (одинаковую фонетическую транскрипцию), но различное написание, их называют **омофонами**, например: *«порог-порок-парок» — «pàrok», «луг-лук» — «luk», «туш-тушь» — «tush», «бал-балл» — «bal».*

Если же слова имеют как одинаковое написание, так и одинаковое произношение, их называют полными **омонимами**, например, слово «бор» может обозначать химический элемент или сосновый лес, слово «эфир» может означать как органическое вещество, так и радиовещание и телевидение.

Кроме того, существует ещё класс слов, называемых **паронимами** — слова, сходные по звучанию (близкие по фонемному составу и акцентной структуре), но разные по значению и строению. Часто ошибочно одно употребляется вместо другого. Например, *«адресат» — «адресант», «экскаватор» — «эскалатор», «абонемент» — «абонент», «экономический» — «экономичный» — «экономный», «ординарный» — «одинарный»*.

Такого рода неоднозначности могут «запутать» систему распознавания, что может привести к нелепой или ошибочной реакции системы. Например, пользователь хочет узнать список лекарств от головной боли: *«Что поможет, если голова болит?»,* однако, система может не выдать правильного ответа, так как будет пытаться обработать запрос *«Что поможет, если голова — болид?».*

Более сложный пример ошибки при паронимии: пользователь, задавая различные вопросы о театре и кино, на очередной вопрос *«Кто такая Сара Бернар?»,* получит неожиданный ответ *«Порода собак по имени Сара. Происходит от азиатских догообразных собак»*, так как система «решила», что пользователь спрашивает о сенбернарах.

Решение такого рода задач, называемых задачами разрешения (снятия) омонимии, во многом удаётся достичь методами статистического анализа текстов. Однако окончательное её решение возможно только с использованием методов семантического анализа, являющихся составной частью систем автоматической обработки текстов.

Системы автоматической обработки текстов (АОТ)

Для решения задач анализа текстов существует множество различных систем, ориентированных под разные задачи и языки. Наиболее известной системой АОТ для английского языка является *Stanford Parser [3].* Для нашей задачи по ряду причин наиболее подходящим выглядит программный комплекс АОТ для русского языка [4], который обеспечивает:

- поддержку русского языка;
- поверхностно-семантический анализ текста;
- открытость исходного кода.

Программный комплекс автоматической обработки текстов АОТ включает в себя следующие компоненты [5]:

- графематический анализ — выделение слов, цифровых комплексов, дат, формул и пр.;
- морфологический анализ — морфологическая интерпретация слов;
- синтаксический анализ — построение дерева синтаксических зависимостей текста;
- семантический анализ — построение поверхностно-семантического графа.

Для каждого компонента комплекса существует свой язык представления данных, состоящий из понятий текущего уровня абстракции структуры текста и правил их комбинирования.

Для решения поставленной задачи разрешения омонимии будут использованы результаты синтаксического и семантического уровней.

Синтаксический анализ — один из этапов обработки текста, задачей которого является построение синтаксических групп на одном морфологическом варианте простого предложения с использованием синтаксических правил. Понятия, используемые в языке представления на синтаксическом уровне разбора текста, включают в себя типы синтаксических фрагментов и синтаксических групп [6].

Типы синтаксических фрагментов включают в себя:

- деепричастный оборот (ДПР);
- причастный оборот (ПРЧ);
- вводный оборот (ВВОД);
- необособленное согласованное определение в препозиции (НСО);
- фрагмент с личной формой глагола (ГЛ_ЛИЧН);
- фрагмент с кратким причастием (КР_ПРЧ);
- фрагмент с кратким прилагательным (КР_ПРИЛ);
- фрагмент с предикативом (ПРЕДК), с инфинитивом (ИНФ);
- фрагмент с тире (ТИРЕ);
- фрагмент со сравнительным прилагательным (СРАВН).

К типам синтаксических групп относятся:

- количественная группа (КОЛИЧ);
- последовательность чисел вперемешку со знаками препинания (КОЛИЧ);
- существительное из заданного перечня и числовой идентификатор (СУЩ-ЧИСЛ);
- слова степени (типа «очень») с группой прилагательного или причастия (НАР_ПРИЛ);
- однородные прилагательные (ОДНОР_ПРИЛ);
- однородные наречия (ОДНОР_НАР);
- однородные инфинитивы (ОДНОР_ИНФ);
- однородные прилагательные сравнительной степени (ОДНОР_ПРИЛ);
- группы даты (ДАТА);
- группа временных отрезков (СЛОЖ_ПГ);
- аналитическая форма сравнительной степени прил. Или наречия (СРАВН-СТЕПЕНЬ);
- наречие и глагол (НАРЕЧ-ГЛАГОЛ);
- одно или несколько прилагательных, согласованных по роду, числу и падежу со стоящим сразу после них существительным (ПРИЛ-СУЩ);
- наречное числительное и именная группа (НАР-ЧИСЛ-СУЩ);
- числительное и именная группа (ЧИСЛ-СУЩ);
- генитивная пара (ГЕНИТ_ИГ);
- предложная группа (ПГ);
- однородные именные группы (ОДНОР_ИГ);
- отрицание и глагольная форма (ОТР_ФОРМА);

- глагольная форма и контактное прямое дополнение (ПРЯМ_ДОП);
- группа электронного адреса (ЭЛ_АДРЕС);
- глагольная форма и контактный инфинитив (ГЛАГ_ИНФ);
- подлежащее (ПОДЛ);
- сказуемое (вершина клаузы).

Семантический анализ — этап обработки текста, задачей которого является построение семантической структуры предложения, состоящей из семантических узлов (понятий) и семантических отношений на этих понятиях.

Каждому семантическому узлу приписан ряд атрибутов:

- набор графематических слов, из которых состоит данный узел;
- номер семантически главного слова в узле;
- грамматическая интерпретация узла (внешняя синтаксическая характеристика);
- номер фрагмента (клаузы), которому принадлежит узел;
- предлог, который в синтаксисе управлял этим узлом;
- ссылка на словарную статью в семантических словарях, которая является интерпретацией этого узла (может быть не определена);
- ссылка на словарную статью открытого словосочетания и номер элемента в поле состав этого словосочетания (может быть не определён) и др.

Параметрами семантического отношения могут быть:

- имя отношения;
- ссылка на словарную статью, откуда было взято это отношение, и номер валентности в этой статье;
- перечень русских слов, которые являются лексическими реализациями этого отношения во входном тексте (предлоги, союзы и т.д.);
- русское синтаксическое отношение, которое является реализацией этого семантического отношения.

В языке представления используются следующие понятия [7]:

• автор(AUTHOR)	• агент (AGENT)
• адресат (ADR)	• в направлении (IN-DIRECT)
• время (TIME)	• значение (VALUE)
• идентификатор (IDENT)	• имя (NAME)
• инструмент (INSTR)	• исходная точка (SRC-PNT)
• контрагент (C-AGENT)	• количество (QUANTIT)
• конечная точка (TRG-PNT)	• локация (LOC)
• масштаб (SCALE)	• материал (MATER)
• назначение (PURP)	• объект (OBJ)
• ограничение (RESTR)	• оценка (ESTIM)
• параметр (PARAM)	• пациент (PACIEN)
• посредник (MEDIATOR)	• признак (PROPERT)
• принадлежность (BELNG)	• причина (CAUSE)
• результат (RESLT)	• содержание (CONTEN)
• способ (METHOD)	• средство (MEANS)
• степень (DEGREE)	• субъект (SUB)
• тема (THEME)	• цель (AIM)

Для каждого синтаксического варианта фрагмента строится множество семантических представлений, лучшее из которых и является результатом работы семантического анализа.

Решение задачи снятия омонимии

Мы ограничимся здесь рассмотрением случаев, когда на вход модуля распознавания речи поступают отдельные речевые команды в виде достаточно простых фраз, состоящих из небольшого количества слов. Обработка начинается с получения нескольких гипотез от модуля распознавания речи. Так как омофоны фонетически неразличимы, то и их вероятности в гипотезах будут равными. В дальнейшем остаётся решить задачу снятия омонимии, т.е. решить, какой из гипотез (распознанных фраз) отдать предпочтение.

Процесс снятия омонимии можно представить как последовательность следующих шагов:

- отсеивание гипотез на основе синтаксического анализа;
- отсеивание гипотез на основе семантического анализа;
- отсеивание гипотез на основе анализа семантической (или статистической) близости словосочетаний.

На первом этапе для каждой из гипотез строится синтаксическое дерево разбора. Далее отсеиваются гипотезы, для которых дерево разбора не было построено, что указывает на синтаксическую некорректность распознанных предложений. Если на этом шаге осталась только одна гипотеза, то дальнейшего анализа не происходит, а оставшаяся гипотеза возвращается как результат. На данном этапе есть возможность снятия омонимии, если омонимы имеют различные морфологические характеристики.

Например, *«лес» — «лез», «старожил» — «сторожил», «течь»* (протекать) *и «течь»* (протекание).

На втором этапе для всех гипотез строится поверхностно-семантический граф. Далее отсеиваем гипотезы, для которых дерево разбора не было построено, что указывает на то, что не все понятия в предложении имеют общие семантические отношения с остальными понятиями в предложении. Если на этом шаге осталась только одна гипотеза, то дальнейшего анализа не происходит, а оставшаяся гипотеза возвращается как результат. На данном этапе есть возможность снятия омонимии, если омонимы имеют различные семантические характеристики, т.е. участвуют в различных семантических отношениях, что значит, имеют различный смысл или домен в иерархии понятий.

Третий этап анализа необходим в тех случаях, когда на предыдущих этапах не удалось снять омонимию. Это может случиться, если более чем для одной гипотезы были построены корректные как синтаксическое, так и семантическое деревья, или противоположный случай, когда ни для одной из гипотез не было построено корректного синтаксического или семантического дерева разбора. В таком случае нельзя однозначно сказать, какая из гипотез верная, но можно сделать предположение, основываясь на следующих признаках:

- близость одного из омонимов к контексту диалога;
- большая статистическая вероятность встречаемости словосочетания в одной из гипотез (N-граммы).

Для учёта контекста необходимо сохранять список значащих понятий последних обработанных предложений. Такой список отражает тема-рематическую [8] последовательность связи нескольких последних предложений диалога. Список понятий контекста диалога должен быть ограничен по длине и построен по принципу очереди: *«первым добавили — последним изъяли»*.

Расчёт близости понятий рассчитывается как кратчайший путь в семантическом графе связи понятий: по ребрам отношений гипоним/гипероним и синоним. Для этого удобно использовать различные тезаурусы, такие как Wiktionary (wiktionary.org), RussNet (project.phil.spbu.ru/) или UNLWEB (http://unlweb.net). Среди гипотез выбирается та, среднее расстояние которой ближе к текущему контексту диалога.

Приведём далее несколько примеров снятия омонимии рассмотренными выше способами.

В том случае, если в ряде омофонов есть морфологические различия, задача разделения омофонов, как уже было сказано, может быть решена уже **на первом этапе** методами синтаксического анализа. Рассмотрим омофоны *«старожил — сторожил»* во фразе *«Где деревенский старожил лет сорок с ключницей бранился»* (А.С. Пушкин). На выходе модуля распознавания речи появилось две гипотезы: *«Где деревенский старожил лет сорок с ключницей бранился»*, *«Где деревенский сторожил лет сорок с ключницей бранился»*. Используя синтаксический анализатор АОТ на этих фразах мы получим результаты, отображённые на *рис. 1*.

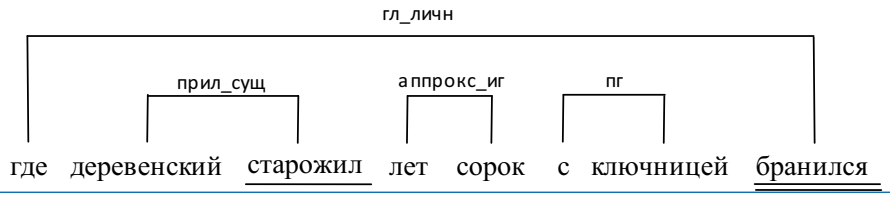

Рис. 1. Дерево синтаксического разбора для фразы: «Где деревенский старожил лет сорок с ключницей бранился»

Из *рис. 1* видно, что полное синтаксическое дерево удалось построить для первого примера и проводить дальнейший семантический анализ нет необходимости.

Рассмотрим более сложный пример омофонов *«балл — бал»* во фразе *«Мы решили поехать на осенний бал»*. На выходе модуля распознавания речи две гипотезы: *«Мы решили поехать на осенний бал»*, *«Мы решили поехать на осенний балл»*. Используя синтаксический анализатор АОТ, для этих фраз мы получим результаты, представленные на *рис. 2.*

Так как омофоны *«балл — бал»* имеют сходные морфологические характеристики, то и деревья разбора предложений получились сходными и полными. В этом случае, исходя из синтаксического анализа, нельзя сделать вывод о достоверности гипотез, и мы переходим к семантическому анализу данных фраз.

Для проведения анализа на втором этапе методами семантического анализа нам необходимо знать семантические характеристики каждого омофона, для этого в программном комплек-

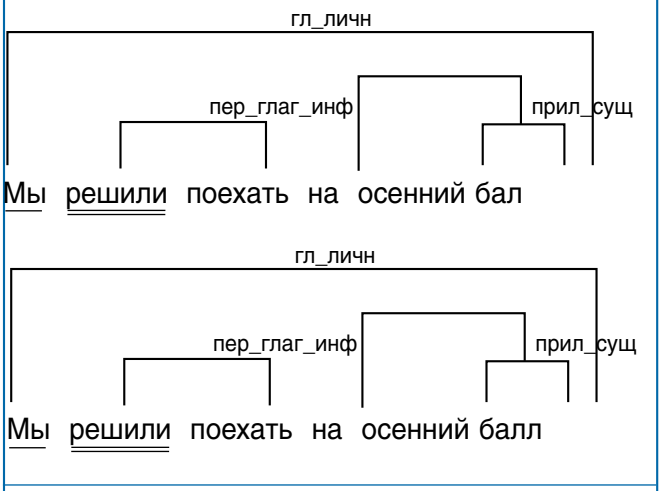

Рис. 2. Деревья синтаксического разбора для гипотез «Мы решили поехать на осенний бал/балл»

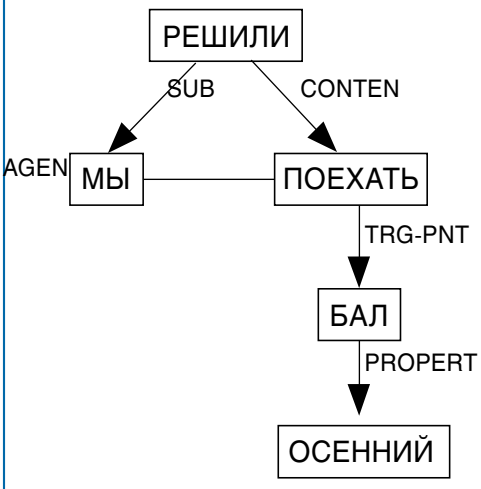

Рис. 3. Дерево семантического разбора для фразы «Мы решили поехать на осенний бал»

се АОТ используется Русский общесемантический словарь (РОСС). В статьях семантического словаря указаны различные семантические отношения, в которых может участвовать описываемое слово (понятие). Используя поверхностно-семантический анализатор АОТ, мы получим граф, представленный на *рис. 3*.

Слова «*поехать*» и «*бал*» связаны отношением «конечный пункт». Для гипотезы «*Мы решили поехать на осенний балл*» построить связный граф нельзя, так как слово «балл» не может участвовать в отношении «конечный пункт». Таким образом, у нас осталась только одна гипотеза, и дальнейшего анализа не требуется.

Рассмотрим пример решения задачи снятия омонимии **на третьем этапе** анализа распознанной фразы. Если омонимы обладают сходными морфологическими характеристиками, их сложно разделить на этапе синтаксического анализа. В том случае если омонимы обладают к тому же и сходными семантическими характеристиками, либо о некоторых омонимах нет информации, то разделить их на этапе семантического анализа также нет возможности. Примерами омонимов со сходными синтаксическими и семантическими характеристиками могут служить: «*пруд*» — «*прут*», «*замо́к*» — «*замо̀к*» и др. Семантическая информация о словах профессиональной лексики, жаргонов и других, редко используемых слов может отсутствовать в семантическом словаре. Пример такого рода омонимов: «*осветить*» — «*освятить*», «*лук*» — «*луг*» и др.

Рассмотрим произнесённую фразу: «*Коля выломал тогда ивовый прут и стукнул по телеге*» (В. Белов) и гипотезы её распознавания: «*Коля выломал тогда ивовый прут и стукнул по телеге*» и «*Коля выломал тогда ивовый пруд и стукнул по телеге*». Синтаксические и семантические деревья разбора, построенные программным комплексом АОТ для этих гипотез, выглядят одинаково и представлены на *рис. 4*.

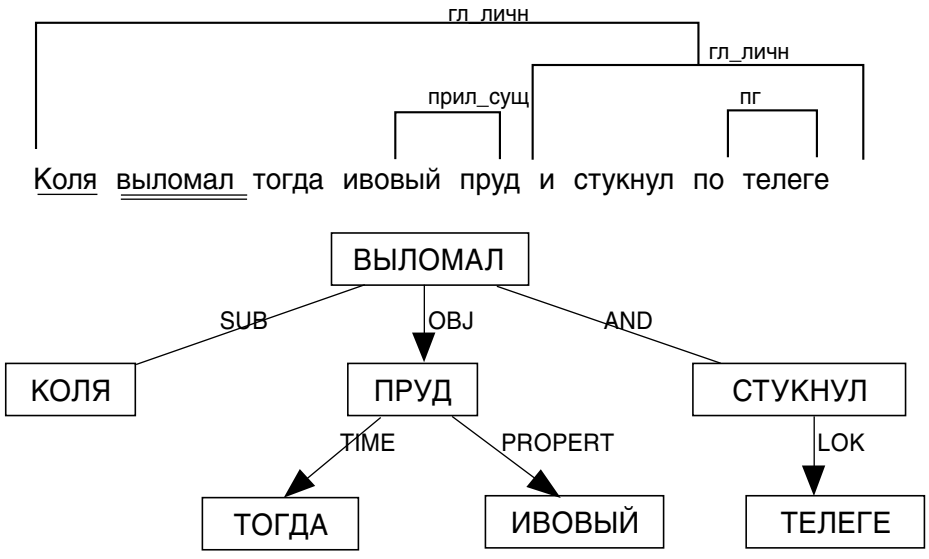

Рис. 4. Деревья синтаксического и семантического разбора для гипотез «Коля выломал тогда ивовый прут/пруд и стукнул по телеге»

Так как синтаксический и семантический этапы не дали результата, переходим на третий этап анализа. Рассчитаем минимальное расстояние для понятий *«пруд»* и *«прут»* с другими понятиями из их отношений (анализ на базе семантического словаря Wiktionary). Для понятия *«пруд»:*

«пруд» — «водоём» — «природные географические объекты» — «география» — «естественные науки» — «ботаника» — «растения» — «ива» — «ивовый»;

итого 8 переходов между понятиями *«пруд»* и *«ивовый»*. Для понятия *«прут»*:

«прут/ветка/побег» — «ботанические термины» — «ботаника» — «растения» — «ива» — «ивовый»;

итого 5 переходов и 2 перехода по синонимам между понятиями *«прут»* и *«ивовый»*. Таким образом, находим, что использование понятия *«прут»* более вероятно в данном контексте фразы.

При рассмотрении гипотез *«Иван осветил фонариком»* и *«Иван освятил фонариком»* обе гипотезы являются грамматически и семантически корректными. Для того чтобы выбрать более правдоподобную из них, необходимо рассчитать расстояние между понятиями *«осветить»* и *«фонарик»* («осветить» — «свет» — «фонарик») и понятиями «освятить» и «фонарик». Очевидно, что расстояние в первом случае будет значительно меньше, что говорит о большей вероятности первой гипотезы.

При практическом использовании систем распознавания речи наиболее часто приходится иметь дело не со словами-омофонами (полностью совпадающими по звучанию, т.е. по фонемному составу), а со словами-паронимами, близкими по фонемному составу и по акцентной структуре. Рассмотрим пример распознавания устной речи системой «СТЕНОГРАФ», построенной на основе интернет-приложения *Google Speech Recognition* [9].

Произнесённая фраза: ***«Они собрали все кости».***
Варианты гипотез системы «СТЕНОГРАФ»:

Они собрались все гости;
Они собрали все гости;
Они собрали все кости;
Они собрали всех кости;
Они собрались все кости;
Они собрание всех кости;
Они собрать и все кости;
Они собрать всех кости;
Они собрали все кости.

В данном примере паронимами являются слова: *«собрали — собрать — собрание — собрались» и «гости — кости»*. Как видно из полученного результата, истинную гипотезу «СТЕНОГРАФ» поместил на последнее место. Это произошло оттого, что в основе стратегии распознавания *Google* лежат статистические методы и фраза *«Они собрались все гости»* оказалась значительно более вероятной, чем *«Они собрали все кости»*.

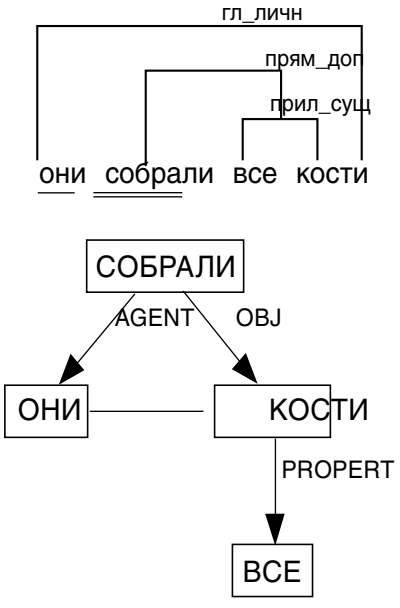

Рис. 5. Деревья синтаксического и семантического разбора для гипотезы «Они собрали все кости»

В описанном эксперименте для произнесённой фразы *«Они собрали все кости»* после синтаксического анализа остались следующие гипотезы: *«Они собрали все кости», «Они собрали всех кости», «Они собрали все гости», «Они собрались все кости», «Они собрались все гости».*

После семантического анализа осталась только одна верная гипотеза: *«Они собрали все кости»*. Синтаксическое и семантическое деревья разбора представлены на *рис. 5*.

Заключение

В статье были рассмотрены основные сценарии решения задачи снятия омонимии при распознавании речи, описаны существующие средства и ресурсы, позволяющие провести синтаксический и семантический анализы текста, показана практическая применимость описываемого метода.

Использование синтаксического анализа гипотез при распознавании речи позволяет отсеять большинство неверных гипотез. Однако так как большое количество фраз с омонимами обладает сходными синтаксическими свойствами, обусловленными в большинстве своём их языковым родством, то после обработки синтаксической структуры неоднозначность не снимается

полностью. Для снятия омонимии в таких случаях необходимо проводить наряду со статистическим N-грамм анализом также и анализ семантической структуры предложения.

Литература

1. Лобанов Б.М. Проблема разрешения «ё»-омографов при синтезе речи по тексту // Труды Международной конференции «Компьютерная лингвистика и интеллектуальные технологии» (Диалог'2009), 1–4 июня 2009. М.: Наука, 2009. С. 330–338.
2. Цирульник Л.И., Барбук С.Г., Лобанов Б.М. Статистический анализ и контекстуальные правила разрешения графической омонимии при синтезе речи по тексту // Труды Международной конференции «Компьютерная лингвистика и интеллектуальные технологии» (Диалог'2009) / Москва, 27–31 мая 2009. С. 530–536.
3. NLP Stanford, 2013. [Electronic resource] Mode of access: http://nlp.stanford.edu — Date of access: 21.11.2013.
4. AOT, 2013. [Electronic resource] Mode of access: http://www.aot.ru — Date of access: 21.11.2013.
5. Sokirko A.A. Short description of Dialing Project // Technical documentation [Electronic resource]. 2013. Mode of access: http://www.aot.ru/docs/sokirko/sokirko-candid-eng.html — Date of access: 21.11.2013.
6. Панкратов Д.В., Гершензон Л.М., Ножов И.М. Описание фрагментации и синтаксического анализа в системе Диалинг // Техническая документация [Electronic resource]. 2000. Mode of access: http://http://www.aot.ru/docs/synan.html_— Date of access: 21.11.2013.
7. Леонтьева Н.Н. Русский общесемантический словарь (РОСС): структура, наполнение. НТИ. Сер. 2. 1997. N 12. С. 5–20.
8. Валгина Н.С. Теория текста: учебное пособие. М.: Изд-во МГУП «Мир книги», 1998.
9. Житко В.А., Лобанов Б.М. Применение облачных интернет-технологий при распознавании речи // Информатика. № 4. Минск. 2012.

Сведения об авторах

Лобанов Борис Мефодьевич,
доктор технических наук, профессор, главный научный сотрудник Лаборатории распознавания и синтеза речи ОИПИ НАН Беларуси, профессор Университета в Белостоке (Польша), основатель и лидер научной школы по проблеме распознавания и синтеза речи. Область научных интересов: теория и методы синтеза, распознавания и понимания устной речи, речевые технологии, компьютерная лингвистика.
E-mail: Lobanov@newman.bas-net.by

Житко Владимир Александрович,
аспирант кафедры интеллектуальных информационных технологий Белорусского государственного университета информатики и радиоэлектроники, г. Минск, Республика Беларусь. Область научных интересов: семантика, диалоговые системы, распознавание речи. E-mail: zhitko.vladimir@gmail.com